

Nonparametric Ensemble Estimation of Distributional Functionals

Kevin R. Moon*, Kumar Sricharan[†], Kristjan Greenewald*, Alfred O. Hero III*

*EECS Dept., University of Michigan, {krmoon,greenewk,hero}@umich.edu

[†]Xerox PARC, sricharan.kumar@parc.com

Abstract

Distributional functionals are integral functionals of one or more probability distributions. Distributional functionals include information measures such as entropy, mutual information, and divergence. Recent work has focused on the problem of nonparametric estimation of entropy and information divergence functionals. Many existing approaches are restrictive in their assumptions on the density support set or require difficult calculations at the support boundary which must be known *a priori*. The MSE convergence rate of a leave-one-out kernel density plug-in divergence functional estimator for general bounded density support sets is derived where knowledge of the support boundary is not required. The theory of optimally weighted ensemble estimation is generalized to derive two estimators that achieve the parametric rate when the densities are sufficiently smooth. The asymptotic distribution of these estimators and some guidelines for tuning parameter selection are provided. Based on the theory, an empirical estimator of Rényi- α divergence is proposed that outperforms the standard kernel density plug-in estimator, especially in high dimension. The estimators are shown to be robust to the choice of tuning parameters.

I. INTRODUCTION

Distributional functionals are integral functionals of one or more probability distributions. The most common distributional functionals in information theory are entropy, mutual information, and information divergence which have many applications in the fields of information theory, statistics, signal processing, and machine learning. Information divergence is the most general of these information measures and is a measure of the difference between probability distributions. Some applications involving divergences include estimating the decay rates of error probabilities [1], estimating bounds on the Bayes error for a classification problem [2]–[8], extending machine learning algorithms to distributional features [9]–[12], testing the hypothesis that two sets of samples come from the same probability distribution [13], clustering [14]–[16], feature selection and classification [17]–[19], blind source separation [20], [21], image segmentation [22]–[24], and steganography [25]. For many more applications of divergence measures, see [26].

Mutual information and entropy are both special cases of divergences and have been used in some of the above applications as well as others such as determining channel capacity [1], fMRI data processing [27], intrinsic dimension estimation [28], [29], and texture classification and image registration [30]. An important subset of information divergences is the family of f -divergences [31], [32]. This family includes the well-known Kullback-Leibler (KL) divergence [33], the Rényi- α divergence [34], the Hellinger-Bhattacharyya distance [35], [36], the Chernoff- α divergence [5], the total variation distance, and the Henze-Penrose divergence [6].

We consider the problem of estimating distributional functionals when only a finite population of independent and identically distributed (i.i.d.) samples is available from each d -dimensional distribution that is unknown, nonparametric, and smooth. For simplicity, we focus on functionals of two distributions, i.e. divergence functionals. However, our methods are general enough that they can be easily extended to estimate functionals of any finite number of distributions. While several estimators of divergence functionals have been previously defined, the convergence rates are known for only a few of them. Furthermore, the asymptotic distributions of these estimators is unknown for nearly all of them. Thus these estimators cannot be easily used to perform inference tasks on the divergence such as testing that two populations have identical distributions or constructing confidence intervals. In this paper, we derive mean squared error (MSE) convergence rates for kernel density plug-in divergence functional estimators. We then generalize the theory of optimally weighted ensemble entropy estimation developed in [37] to obtain two divergence functional estimators with a MSE convergence rate of $O\left(\frac{1}{N}\right)$, where N is the sample size, when the densities are sufficiently smooth. We obtain the asymptotic distribution of the weighted ensemble estimators which enables us to perform hypothesis testing. We then examine the problem of tuning parameter selection. Finally, we empirically validate the theory and establish the estimators' robustness to the choice of tuning parameters.

A. Related Work

Several nonparametric estimators for some functionals of two distributions including f -divergences already exist. For example, Póczos and Schneider [9] established weak consistency of a bias-corrected k -nn estimator for Rényi- α and other divergences

This work was partially supported by ARO MURI grant W911NF-15-1-0479, NSF grant CCF-1217880, and a NSF Graduate Research Fellowship to the first author under Grant No. F031543.

of a similar form where k is fixed. Wang et al [38] provided a k -nn based estimator for the KL divergence. Mutual information and divergence estimators based on plug-in histogram schemes have been proven to be consistent [39]–[42]. Hero et al [30] provided an estimator for Rényi- α divergence but assumed that one of the densities was known. However none of these works study the convergence rates nor the asymptotic distribution of their estimators.

There has been recent interest in deriving convergence rates for divergence estimators [43]–[48]. The rates are typically derived in terms of a smoothness condition on the densities, such as the Hölder condition [49]:

Definition 1 (Hölder Class): Let $\mathcal{X} \subset \mathbb{R}^d$ be a compact space. For $r = (r_1, \dots, r_d)$, $r_i \in \mathbb{N}$, define $|r| = \sum_{i=1}^d r_i$ and $D^r = \frac{\partial^{|r|}}{\partial x_1^{r_1} \dots \partial x_d^{r_d}}$. The Hölder class $\Sigma(s, K)$ of functions on $L_2(\mathcal{X})$ consists of the functions f that satisfy

$$|D^r f(x) - D^r f(y)| \leq K \|x - y\|^{s-r},$$

for all $x, y \in \mathcal{X}$ and for all r s.t. $|r| \leq \lfloor s \rfloor$.

From Definition 1, it is clear that if a function f belongs to $\Sigma(s, K)$, then f is continuously differentiable up to order $\lfloor s \rfloor$. In this work, we propose estimators that achieve the parametric MSE convergence rate of $O(1/T)$ when $s \geq d$ and $s \geq \frac{d+1}{2}$, respectively.

Nguyen et al [44] proposed a method for estimating f -divergences by estimating the likelihood ratio of the two densities by solving a convex optimization problem and then plugging it into the divergence formulas. For this method they prove that the minimax MSE convergence rate is parametric ($O(\frac{1}{T})$) when the likelihood ratio is in the bounded Hölder class $\Sigma(s, K)$ with $s \geq d/2$. However, this estimator is restricted to true f -divergences and may not apply to the broader class of divergence functionals (as an example, the L_2^2 divergence is not an f -divergence). Additionally, solving the convex problem of [44] has similar complexity to that of training a support vector machine (SVM) (between $O(T^2)$ and $O(T^3)$) which can be demanding when T is very large. In contrast, our method of optimally weighted ensemble estimation depends only on simple density plug-in estimates and the solution of an offline convex optimization problem. Thus the most computationally demanding step in our approach is the calculation of the density estimates, which has complexity no greater than $O(T^2)$.

Singh and Póczos [46], [47] provided an estimator for Rényi- α divergences as well as general density functionals that uses a “mirror image” kernel density estimator. They prove a convergence rate of $O(\frac{1}{T})$ when $s \geq d$ for each of the densities. However this method requires several computations at each boundary of the support of the densities which becomes difficult to implement as d gets large. Also, this method requires knowledge of the support of the densities which may not be possible for some problems. In contrast, while our assumptions require the density supports to be bounded, knowledge of the support is not required for implementation.

The “linear” and “quadratic” estimators presented by Krishnamurthy et al [45] estimate divergence functionals that include the form $\int f_1^\alpha(x) f_2^\beta(x) d\mu(x)$ for given α and β where f_1 and f_2 are probability densities. These estimators achieve the parametric rate when $s \geq d/2$ and $s \geq d/4$ for the linear and quadratic estimators, respectively. However, the latter estimator is computationally infeasible for most functionals and the former requires numerical integration for some divergence functionals, which can be computationally difficult. Additionally, while a suitable α - β indexed sequence of divergence functionals of this form can be made to converge to the KL divergence, this does not guarantee convergence of the corresponding sequence of divergence estimators in [45], whereas our estimator can be used to estimate the KL divergence. Other important f -divergence functionals are also excluded from this form including some that bound the Bayes error [2], [4], [6]. In contrast, our method applies to a large class of divergence functionals and avoids numerical integration.

Finally, Kandasamy et al [48] propose influence function based estimators of distributional functionals that achieve the parametric rate when $s \geq d/2$. While their method can be applied to general functionals, their estimator requires numerical integration for some functionals. Additionally, the estimators in both Kandasamy et al [48] and Krishnamurthy et al [45] require an optimal kernel density estimator. This is difficult to construct when the density support is bounded as it requires knowledge of the support boundary and difficult computations at the boundary, whereas our method does not require knowledge of the support boundary.

Asymptotic normality has been established for certain appropriately normalized divergences between a specific density estimator and the true density [50]–[52]. This differs from our setting where we assume that both densities are unknown. The asymptotic distributions of the estimators in [44]–[47] are currently unknown. Kandasamy et al [48] prove a central limit theorem for their data-splitting estimator but do not prove similar results for their leave-one-out estimator. We establish a central limit theorem for our proposed leave-one-out divergence estimator.

Divergence functional estimation is also related to the problem of entropy functional estimation which has received a lot of attention. Some examples include [53]–[55] which used specialized kernel density estimators to achieve the parametric convergence rate when the density has smoothness parameter $s \geq d/4$. Sricharan et al [37] derived an entropy functional estimator that uses a weighted average of an ensemble of bias-corrected estimators. While the approach in [37] requires the density to have smoothness parameter $s \geq d$ in order to achieve the parametric rate, their approach is simpler to implement compared to the estimators in [53]–[55] as long as the density support set is known.

This paper extends the work in [37] to functionals of two distributions and improves upon the results reported in [7], [43] which explored k -nearest neighbor (k -nn) based estimators of f -divergences. We derive more general conditions on the required smoothness of the densities for the MSE convergence rates of plug-in kernel density estimators (KDE) of general divergence functionals. We then generalize the theory of optimally weighted ensemble entropy estimation developed in [37] to obtain two divergence functional estimators that achieve the parametric MSE convergence rate of $O(\frac{1}{N})$ when the densities are sufficiently smooth, where N is the sample size. One of these estimators achieves this rate when the densities have $s \geq (d+1)/2$, whereas [43] requires $s \geq d$. These estimators apply to general divergence functionals and are simpler to implement than other estimators that also achieve the parametric rate. This work also extends these estimators to more general bounded density support sets in \mathbb{R}^d , whereas the proofs in [43] restricted the estimator to compactly supported densities with no boundary conditions (e.g., a support set equal to the surface of a torus), which is unnecessarily restrictive. Finally, we use a leave-one-out approach that uses all of the data for both density estimation and integral approximation in contrast to [7], [37], [43], [56] and others which use a less efficient data-splitting approach. We then derive the asymptotic distribution of the weighted ensemble estimators which enables us to construct confidence intervals and perform hypothesis testing.

B. Organization and Notation

The paper is organized as follows. Section II presents the kernel density plug-in divergence functional estimator and its MSE convergence rate. Our generalized theory of optimally weighted ensemble estimation and the proposed ensemble estimators are given in Section III. A central limit theorem for the ensemble estimators is also presented in Section III. In Section IV, we provide guidelines for selecting the tuning parameters based on experiments and the theory derived in the previous sections. We then perform experiments in Section IV that validate the theory and establish the robustness of the proposed estimators to the tuning parameters.

Bold face type is used for random variables and random vectors. The conditional expectation given a random variable \mathbf{Z} is denoted $\mathbb{E}_{\mathbf{Z}}$. The variance of a random variable is denoted \mathbb{V} and the bias of an estimator is denoted \mathbb{B} .

II. THE DIVERGENCE FUNCTIONAL WEAK ESTIMATOR

This paper focuses on estimating functionals of the form

$$G(f_1, f_2) = \int g(f_1(x), f_2(x)) f_2(x) dx, \quad (1)$$

where $g(x, y)$ is a smooth functional, and f_1 and f_2 are smooth d -dimensional probability densities. If $g(f_1(x), f_2(x)) = g\left(\frac{f_1(x)}{f_2(x)}\right)$, g is convex, and $g(1) = 0$, then $G(f_1, f_2)$ defines the family of f -divergences. Some common divergences that belong to this family include the KL divergence ($g(t) = -\ln t$), the Rényi- α divergence ($g(t) = t^\alpha$), and the total variation distance ($g(t) = |t - 1|$). In this work, we consider a broader class of functionals than the f -divergences.

A. The Kernel Density Plug-in Estimator

We use a kernel density plug-in estimator of the divergence functional in (1). Assume that N_1 i.i.d. realizations $\{\mathbf{Y}_1, \dots, \mathbf{Y}_{N_1}\}$ are available from f_1 and N_2 i.i.d. realizations $\{\mathbf{X}_1, \dots, \mathbf{X}_{N_2}\}$ are available from f_2 . Let $h_i > 0$ be the kernel bandwidth for the density estimator of f_i . Let $K(\cdot)$ be a kernel function with $\|K\|_\infty < \infty$ where $\|K\|_\infty$ is the ℓ_∞ norm of the kernel K . The KDEs are

$$\begin{aligned} \tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j) &= \frac{1}{N_1 h_1^d} \sum_{i=1}^{N_1} K\left(\frac{\mathbf{X}_j - \mathbf{Y}_i}{h_1}\right), \\ \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j) &= \frac{1}{M_2 h_2^d} \sum_{\substack{i=1 \\ i \neq j}}^{N_2} K\left(\frac{\mathbf{X}_j - \mathbf{X}_i}{h_2}\right), \end{aligned}$$

where $M_2 = N_2 - 1$. $G(f_1, f_2)$ is then approximated as

$$\tilde{\mathbf{G}}_{h_1, h_2} = \frac{1}{N_2} \sum_{i=1}^{N_2} g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_i), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_i)\right). \quad (2)$$

B. Convergence Rates

Similar to [7], [37], [43], the principal assumptions we make on the densities f_1 and f_2 and the functional g are that: 1) f_1 , f_2 , and g are smooth; 2) f_1 and f_2 have common bounded support sets \mathcal{S} ; 3) f_1 and f_2 are strictly lower bounded on \mathcal{S} . We also assume 4) that the support is smooth with respect to the kernel $K(u)$. Our full assumptions are:

- (A.0): Assume that the kernel K is symmetric, is a product kernel, and has bounded support in each dimension. Also assume that it has order ν which means that the j th moment of the kernel K_i defined as $\int t^j K_i(t) dt$ is zero for all $j = 1, \dots, \nu - 1$ and $i = 1, \dots, d$ where K_i is the kernel in the i th coordinate.
- (A.1): Assume there exist constants $\epsilon_0, \epsilon_\infty$ such that $0 < \epsilon_0 \leq f_i(x) \leq \epsilon_\infty < \infty, \forall x \in \mathcal{S}$.
- (A.2): Assume that the densities $f_i \in \Sigma(s, K)$ in the interior of \mathcal{S} with $s \geq 2$.
- (A.3): Assume that g has an infinite number of mixed derivatives.
- (A.4): Assume that $\left| \frac{\partial^{k+l} g(x, y)}{\partial x^k \partial y^l} \right|, k, l = 0, 1, \dots$ are strictly upper bounded for $\epsilon_0 \leq x, y \leq \epsilon_\infty$.
- (A.5): Assume the following boundary smoothness condition: Let $p_x(u) : \mathbb{R}^d \rightarrow \mathbb{R}$ be a polynomial in u of order $q \leq r = \lfloor s \rfloor$ whose coefficients are a function of x and are $r - q$ times differentiable. Then assume that

$$\int_{x \in \mathcal{S}} \left(\int_{u: K(u) > 0, x+uh \notin \mathcal{S}} K(u) p_x(u) du \right)^t dx = v_t(h),$$

where $v_t(h)$ admits the expansion

$$v_t(h) = \sum_{i=1}^{r-q} e_{i,q,t} h^i + o(h^{r-q}),$$

for some constants $e_{i,q,t}$.

We focus on finite support kernels for simplicity in the proofs although it is likely that our results extend to some infinitely supported kernels as well. The smoothness assumptions on the densities are weaker as compared to [7], [37], [43]. However, we assume stronger conditions on the smoothness of g to enable us to achieve good convergence rates without knowledge of the boundary of the support set. Assumption A.5 requires the boundary of the density support set to be smooth wrt the kernel $K(u)$ in the sense that the expectation of the area outside of \mathcal{S} wrt any random variable u with smooth distribution is a smooth function of the bandwidth h . It is not necessary for the boundary of \mathcal{S} to have smooth contours with no edges or corners as this assumption is satisfied by the following case:

Theorem 1: Assumption A.5 is satisfied when $\mathcal{S} = [-1, 1]^d$ and when K is the uniform rectangular kernel; that is $K(x) = 1$ for all $x : \|x\|_1 \leq 1/2$.

The proof is given in Appendix A. Given the simple nature of this density support set and kernel, it is likely that other kernels and supports will satisfy A.5 as well. This is left for future work.

Densities for which assumptions A.1 – A.2 hold include the truncated Gaussian distribution and the Beta distribution on the unit cube. Functions for which assumptions A.3 – A.4 hold include $g(x, y) = -\ln\left(\frac{x}{y}\right)$ and $g(x, y) = \left(\frac{x}{y}\right)^\alpha$. The following theorem on the bias follows under assumptions A.0 – A.5:

Theorem 2:

For general g , the bias of the plug-in estimator $\tilde{\mathbf{G}}_{h_1, h_2}$ is of the form

$$\begin{aligned} \mathbb{B}[\tilde{\mathbf{G}}_{h_1, h_2}] &= \sum_{j=1}^r \left(c_{4,1,j} h_1^j + c_{4,2,j} h_2^j \right) + \sum_{j=1}^r \sum_{i=1}^r c_{5,i,j} h_1^j h_2^i + O(h_1^s + h_2^s) \\ &\quad + c_{9,1} \frac{1}{N_1 h_1^d} + c_{9,2} \frac{1}{N_2 h_2^d} + o\left(\frac{1}{N_1 h_1^d} + \frac{1}{N_2 h_2^d} \right). \end{aligned} \quad (3)$$

Furthermore, if $g(x, y)$ has k, l -th order mixed derivatives $\frac{\partial^{k+l} g(x, y)}{\partial x^k \partial y^l}$ that depend on x, y only through $x^\alpha y^\beta$ for some $\alpha, \beta \in \mathbb{R}$,

then for any positive integer $\lambda \geq 2$, the bias is of the form

$$\begin{aligned} \mathbb{B} \left[\tilde{\mathbf{G}}_{h_1, h_2} \right] &= \sum_{j=1}^r \left(c_{4,1,j} h_1^j + c_{4,2,j} h_2^j \right) + \sum_{j=1}^r \sum_{i=1}^r c_{5,i,j} h_1^j h_2^i + O(h_1^s + h_2^s) \\ &\quad + \sum_{j=1}^{\lambda/2} \sum_{m=0}^r \left(c_{9,1,j,m} \frac{h_1^m}{(N_1 h_1^d)^j} + c_{9,2,j,m} \frac{h_2^m}{(N_2 h_2^d)^j} \right) \\ &\quad + \sum_{j=1}^{\lambda/2} \sum_{m=0}^r \sum_{i=1}^{\lambda/2} \sum_{n=0}^r c_{9,j,i,m,n} \frac{h_1^m h_2^n}{(N_1 h_1^d)^j (N_2 h_2^d)^i} \\ &\quad + O \left(\frac{1}{(N_1 h_1^d)^{\frac{\lambda}{2}}} + \frac{1}{(N_2 h_2^d)^{\frac{\lambda}{2}}} \right). \end{aligned} \quad (4)$$

Divergence functionals that satisfy the mixed derivatives condition required for (4) include the KL divergence and the Rényi- α divergence. Obtaining similar terms for other divergence functionals requires us to separate the dependence on h_i of the derivatives of g evaluated at $\mathbb{E}_{\mathbf{Z}} \tilde{\mathbf{f}}_{i,h_i}(\mathbf{Z})$. This is left for future work. See Appendix B for details.

The following variance result requires much less strict assumptions:

Theorem 3: Assume that the functional g in (1) is Lipschitz continuous in both of its arguments with Lipschitz constant C_g . Then the variance of the plug-in estimator $\tilde{\mathbf{G}}_{h_1, h_2}$ is bounded by

$$\mathbb{V} \left[\tilde{\mathbf{G}}_{h_1, h_2} \right] \leq C_g^2 \|K\|_\infty^2 \left(\frac{10}{N_2} + \frac{N_1}{N_2^2} \right).$$

From Theorems 2 and 3, it is clear that we require $h_i \rightarrow 0$ and $N_i h_i^d \rightarrow \infty$ for $\tilde{\mathbf{G}}_{h_1, h_2}$ to be unbiased while the variance of the plug-in estimator depends primarily on the number of samples. Note that the constants in front of the terms that depend on h_i and N_i may not be identical for different i, j, m, n in (3) and (4). However, these constants depend on the densities f_1 and f_2 and their derivatives which are often unknown. The rates given in Thm. 2 and 3 are similar to the rates derived for the entropy plug-in estimator in [37] if $h_i^d = k_i/N_i$. The differences lie in the constants in front of the rates and the dependence on the number of samples from two distributions instead of one. Additionally, as compared to (3), in (4) there are many more terms. These terms enable us to achieve the parametric MSE convergence rate when $s \geq (d+1)/2$ for an appropriate choice of bandwidths whereas the terms in (3) require $s \geq d$ to achieve the same rate.

C. Optimal MSE Rate

From Theorem 2, the dominating terms in the bias are $\Theta(h_i)$ and $\Theta\left(\frac{1}{N_i h_i^d}\right)$. If no attempt is made to correct the bias, the optimal choice of h_i in terms of minimizing the MSE is

$$h_i^* = \Theta\left(N_i^{-\frac{1}{d+1}}\right).$$

This results in a dominant bias term of order $\Theta\left(N_i^{-\frac{1}{d+1}}\right)$. Note that this differs from the standard result for the optimal KDE bandwidth for minimum MSE density estimation which is $\Theta(N^{-1/(d+4)})$ for a symmetric uniform kernel [57].

Figure 1 gives a heatmap showing the leading term $O(h)$ as a function of d and N when $h = N^{-\frac{1}{d+1}}$. The heatmap indicates that the bias of the plug-in estimator in (2) is small only for relatively small values of d .

D. Proof Sketches of Theorems 2 and 3

To prove the expressions for the bias, the bias is first decomposed into two parts by adding and subtracting $g\left(\mathbb{E}_{\mathbf{Z}} \tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}), \mathbb{E}_{\mathbf{Z}} \tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})\right)$ within the expectation creating a “bias” term and a “variance” term. Applying a Taylor series expansion on the bias and variance terms results in expressions that depend on powers of $\mathbb{B}_{\mathbf{Z}} \left[\tilde{\mathbf{f}}_{i,h_i}(\mathbf{Z}) \right] := \mathbb{E}_{\mathbf{Z}} \tilde{\mathbf{f}}_{i,h_i}(\mathbf{Z}) - f_i(\mathbf{Z})$ and $\tilde{\mathbf{e}}_{i,h_i}(\mathbf{Z}) := \tilde{\mathbf{f}}_{i,h_i}(\mathbf{Z}) - \mathbb{E}_{\mathbf{Z}} \tilde{\mathbf{f}}_{i,h_i}(\mathbf{Z})$, respectively. Within the interior of the support, moment bounds can be derived from properties of the KDEs and a Taylor series expansion of the densities. Near the boundary of the support, the smoothness assumption on the boundary A.5 is also required. Note that this approach differs from that in [37] which corrected the KDEs near the boundary of the support set and also used concentration inequalities for the KDEs. The full proof of Thm. 2 is given in Appendix B.

The proof of the variance result takes a different approach. It uses the Efron-Stein inequality which bounds the variance by analyzing the expected squared difference between the plug-in estimator when one sample is allowed to differ. The full proof of Thm. 3 is given in Appendix C.

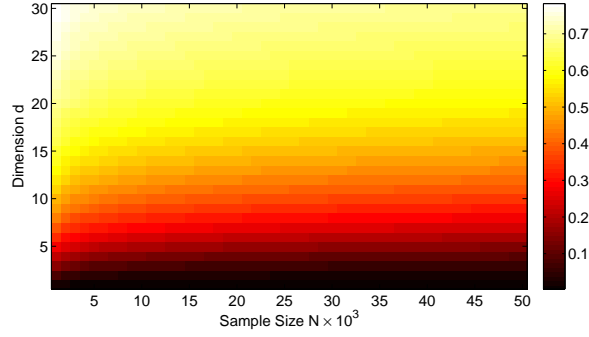


Figure 1. Heat map of predicted bias of divergence functional plug-in estimator based on Theorem 2 as a function of dimension and sample size when $h = N^{\frac{1}{d+1}}$. Note the phase transition in the bias as dimension d increases for fixed sample size N : bias remains small only for relatively small values of d . The proposed weighted ensemble estimator removes this phase transition when the densities are sufficiently smooth.

III. WEIGHTED ENSEMBLE ESTIMATION

As pointed out in Sec. II-C, Thm. 2 shows that when the dimension of the data is not small, the bias of the MSE-optimal plug-in estimator $\tilde{\mathbf{G}}_{h_1, h_2}$ decreases very slowly as a function of sample size, resulting in large MSE. However, by applying the theory of optimally weighted ensemble estimation, originally developed in [37] for entropy estimation, we can modify the minimum MSE estimator by taking a weighted sum of an ensemble of estimators where the weights are chosen to significantly reduce the bias.

A. The Weighted Ensemble Estimator

The bias expression in Theorem 2 is quite complicated due to its dependence on the sample size of two different distributions. We can simplify it significantly by assuming that $N_1 = N_2 = N$ and $h_1 = h_2 = h$. Define $\tilde{\mathbf{G}}_h := \tilde{\mathbf{G}}_{h, h}$.

Corollary 1: For general g , the bias of the plug-in estimator $\tilde{\mathbf{G}}_h$ is given by

$$\mathbb{B}[\tilde{\mathbf{G}}_h] = \sum_{j=1}^{\lfloor s \rfloor} c_{10,j} h^j + c_{11} \frac{1}{N h^d} + O\left(h^s + \frac{1}{N h^d}\right).$$

If $g(x, y)$ has k, l -th order mixed derivatives $\frac{\partial^{k+l} g(x, y)}{\partial x^k \partial y^l}$ that depend on x, y only through $x^\alpha y^\beta$ for some $\alpha, \beta \in \mathbb{R}$, then for any positive integer $\lambda \geq 2$, the bias is

$$\begin{aligned} \mathbb{B}[\tilde{\mathbf{G}}_h] &= \sum_{j=1}^{\lfloor s \rfloor} c_{10,j} h^j + \sum_{q=1}^{\lambda/2} \sum_{j=0}^{\lfloor s \rfloor} c_{11,q,j} \frac{h^j}{(N h^d)^q} \\ &\quad + O\left(h^s + \frac{1}{(N h^d)^{\frac{\lambda}{2}}}\right). \end{aligned}$$

Note that the corollary still holds if N_1 and N_2 are linearly related, i.e., $N = N_1 = \Theta(N_2)$ and similarly if h_1 and h_2 are linearly related, i.e., $h = h_1 = \Theta(h_2)$. We form an ensemble of estimators by choosing different values of h . Choose $\mathcal{L} = \{l_1, \dots, l_L\}$ to be real positive numbers that index $h(l_i)$. Thus the parameter l indexes over different neighborhood sizes for the kernel density estimates. Define $w := \{w(l_1), \dots, w(l_L)\}$ and $\tilde{\mathbf{G}}_w := \sum_{l \in \mathcal{L}} w(l) \tilde{\mathbf{G}}_{h(l)}$. The key to reducing the MSE is to choose the weight vector w to reduce the lower order terms in the bias without substantially increasing the variance.

B. Finding the Optimal Weight

The theory of optimally weighted ensemble estimation is a general theory originally presented by Sricharan et al [37] that can be applied to many estimation problems as long as the bias and variance of the estimator can be expressed in a specific way. We generalize the conditions given in [37] that were required to apply the theory. Let $\mathcal{L} = \{l_1, \dots, l_L\}$ be a set of index values and let N be the number of samples available. For an indexed ensemble of estimators $\{\hat{\mathbf{E}}_l\}_{l \in \mathcal{L}}$ of a parameter E , the weighted ensemble estimator with weights $w = \{w(l_1), \dots, w(l_L)\}$ satisfying $\sum_{l \in \mathcal{L}} w(l) = 1$ is defined as

$$\hat{\mathbf{E}}_w = \sum_{l \in \mathcal{L}} w(l) \hat{\mathbf{E}}_l.$$

$\hat{\mathbf{E}}_w$ is asymptotically unbiased if the estimators $\{\hat{\mathbf{E}}_l\}_{l \in \mathcal{L}}$ are asymptotically unbiased. Consider the following conditions on $\{\hat{\mathbf{E}}_l\}_{l \in \mathcal{L}}$:

- C.1 The bias is expressible as

$$\mathbb{B}[\hat{\mathbf{E}}_l] = \sum_{i \in J} c_i \psi_i(l) \phi_{i,d}(N) + O\left(\frac{1}{\sqrt{N}}\right),$$

where c_i are constants depending on the underlying density, $J = \{i_1, \dots, i_I\}$ is a finite index set with $I < L$, and $\psi_i(l)$ are basis functions depending only on the parameter l and not on the sample size.

- C.2 The variance is expressible as

$$\mathbb{V}[\hat{\mathbf{E}}_l] = c_v \left(\frac{1}{N}\right) + o\left(\frac{1}{N}\right).$$

Theorem 4: Assume conditions C.1 and C.2 hold for an ensemble of estimators $\{\hat{\mathbf{E}}_l\}_{l \in \mathcal{L}}$. Then there exists a weight vector w_0 such that the MSE of the weighted ensemble estimator attains the parametric rate of convergence:

$$\mathbb{E}\left[\left(\hat{\mathbf{E}}_{w_0} - E\right)^2\right] = O\left(\frac{1}{N}\right).$$

The weight vector w_0 is the solution to the following convex optimization problem:

$$\begin{aligned} & \min_w \quad \|w\|_2 \\ & \text{subject to} \quad \sum_{l \in \mathcal{L}} w(l) = 1, \\ & \quad \quad \quad \gamma_w(i) = \sum_{l \in \mathcal{L}} w(l) \psi_i(l) = 0, \quad i \in J. \end{aligned} \tag{5}$$

A more restrictive version of Theorem 4 was originally presented in [37] with the stricter condition of $\phi_{i,d}(N) = N^{-1/(2d)}$. The proof of our generalized version (Theorem 4) is sketched below.

Proof: From condition C.1, the bias of the weighted estimator is

$$\mathbb{B}[\hat{\mathbf{E}}_w] = \sum_{i \in J} c_i \gamma_w(i) \phi_{i,d}(N) + O\left(\frac{\sqrt{L}\|w\|_2}{\sqrt{N}}\right).$$

The variance of the weighted estimator is bounded as

$$\mathbb{V}[\hat{\mathbf{E}}_w] \leq \frac{L\|w\|_2^2}{N}. \tag{6}$$

The optimization problem in (5) zeroes out the lower-order bias terms and limits the ℓ_2 norm of the weight vector w to limit the variance contribution. This results in an MSE rate of $O(1/N)$ when the dimension d is fixed and when L is fixed independently of the sample size N . Furthermore, a solution to (5) is guaranteed to exist as long as $L > I$ and the vectors $a_i = [\psi_i(l_1), \dots, \psi_i(l_L)]$ are linearly independent. This completes our sketch of the proof of Thm. 4. ■

C. Optimally Weighted Distributional Functional (ODin) Estimators

To achieve the parametric rate $O(1/N)$ in MSE convergence it is not necessary that $\gamma_w(i) = 0, i \in J$. Solving the following convex optimization problem in place of the optimization problem in Theorem 4 retains the $O(1/N)$ rate:

$$\begin{aligned} & \min_w \quad \epsilon \\ & \text{subject to} \quad \sum_{l \in \mathcal{L}} w(l) = 1, \\ & \quad \quad \quad \left| \gamma_w(i) N^{\frac{1}{2}} \phi_{i,d}(N) \right| \leq \epsilon, \quad i \in J, \\ & \quad \quad \quad \|w\|_2^2 \leq \eta, \end{aligned} \tag{7}$$

where the parameter η is chosen to achieve a trade-off between bias and variance. Instead of forcing $\gamma_w(i) = 0$, the relaxed optimization problem uses the weights to decrease the bias terms at the rate of $O(1/\sqrt{N})$ yielding an MSE of $O(1/N)$.

We refer to the distributional functional estimators obtained using this theory as **Optimally Weighted Distributional Functional** (ODin) estimators. Sricharan et al [37] applied the stricter version of Theorem 4 to obtain an entropy estimator with convergence rate $O(1/N)$. We also apply the same theory to obtain a divergence functional estimator with the same asymptotic rate. Let $h(l) = lN^{-1/(2d)}$. From Theorem 2, we get $\psi_i(l) = l^i, i = 1, \dots, d$. Note that if $s \geq d$, then we are left with $O\left(\frac{1}{l^d \sqrt{N}}\right)$ in addition to the terms in the sum. To obtain a uniform bound on the bias with respect to w and \mathcal{L} , we also include the function $\psi_{d+1}(l) = l^{-d}$ in the optimization problem. The bias of the resulting base estimator satisfies condition C.1 with $\phi_{i,d}(N) = N^{-i/(2d)}$ for $i = 1, \dots, d$ and $\phi_{d+1,d}(N) = N^{-1/2}$. The variance also satisfies condition C.2. The optimal weight

Algorithm 1 Optimally weighted ensemble estimator of divergence functionals

Input: η , L positive real numbers \mathcal{L} , samples $\{\mathbf{Y}_1, \dots, \mathbf{Y}_N\}$ from f_1 , samples $\{\mathbf{X}_1, \dots, \mathbf{X}_N\}$ from f_2 , dimension d , function g , kernel K

Output: The optimally weighted divergence estimator $\tilde{\mathbf{G}}_{w_0,2}$

- 1: Solve for w_0 using (7) with $\phi_{j,q,d}(N) = N^{-\frac{j+q}{d+1}}$ and basis functions $\psi_{j,q}(l) = l^{j-dq}$, $l \in \bar{l}$, and $\{i, j\} \in J$ defined in (8)
 - 2: **for all** $l \in \bar{l}$ **do**
 - 3: $h(l) \leftarrow lN^{\frac{1}{d+1}}$
 - 4: **for** $i = 1$ to N **do**
 - 5: $\tilde{\mathbf{f}}_{1,h(l)}(\mathbf{X}_i) \leftarrow \frac{1}{Nh(l)^d} \sum_{j=1}^N K\left(\frac{\mathbf{X}_i - \mathbf{Y}_j}{h(l)}\right)$, $\tilde{\mathbf{f}}_{2,h(l)}(\mathbf{X}_i) \leftarrow \frac{1}{(N-1)h(l)^d} \sum_{j=1, j \neq i}^N K\left(\frac{\mathbf{X}_i - \mathbf{X}_j}{h(l)}\right)$
 - 6: **end for**
 - 7: $\tilde{\mathbf{G}}_{h(l)} \leftarrow \frac{1}{N} \sum_{i=1}^N g\left(\tilde{\mathbf{f}}_{1,h(l)}(\mathbf{X}_i), \tilde{\mathbf{f}}_{2,h(l)}(\mathbf{X}_i)\right)$
 - 8: **end for**
 - 9: $\tilde{\mathbf{G}}_{w_0,2} \leftarrow \sum_{l \in \mathcal{L}} w_0(l) \tilde{\mathbf{G}}_{h(l)}$
-

w_0 is found by using (7) to obtain a plug-in divergence functional estimator $\tilde{\mathbf{G}}_{w_0,1}$ with an MSE convergence rate of $O(\frac{1}{N})$ as long as $s \geq d$. Otherwise, if $s < d$ we can only guarantee the MSE rate up to $O(\frac{1}{N^{s/d}})$. We refer to this estimator as the ODin1 estimator.

Another weighted ensemble estimator can be defined that requires less strict assumptions on the smoothness of the densities. This is accomplished by letting $h(l)$ decrease at a faster rate. Let $h(l) = lN^{\frac{1}{d+1}}$. From Theorem 2, we have that if $g(x, y)$ has mixed derivatives of the form of $x^\alpha y^\beta$, then the bias has terms proportional to $l^{j-dq} N^{-\frac{j+q}{d+1}}$ where $j, q \geq 0$ and $j + q > 0$. Theorem 4 can be applied to the ensemble of estimators to derive an estimator that achieves the parametric convergence rate under these conditions. Let $\phi_{j,q,d}(N) = N^{-\frac{j+q}{d+1}}$ and $\psi_{j,q}(l) = l^{j-dq}$. Let

$$J = \{\{j, q\} : 0 < j + q < (d + 1)/2, q \in \{0, 1, 2, \dots, \lfloor d/2 \rfloor\}, j \in \{0, 1, 2, \dots, \lfloor s \rfloor\}\}. \quad (8)$$

Then from (4), the bias of $\tilde{\mathbf{G}}_{h(l)}$ satisfies condition C.1. If $L > |J| = I$, then Theorem 4 can be applied to obtain the optimal weight vector. The estimator $\tilde{\mathbf{G}}_{w_0,2} = \sum_{l \in \mathcal{L}} w_0(l) \tilde{\mathbf{G}}_{h(l)}$ achieves the parametric convergence rate if $s \geq d + 1$ and if $s \geq (d + 1)/2$. Otherwise, if $s < (d + 1)/2$ we can only guarantee the MSE rate up to $O(\frac{1}{N^{2s/(d+1)}})$. $\tilde{\mathbf{G}}_{w_0,2}$ is referred to as the ODin2 estimator and is summarized in Algorithm 1.

D. Comparison of ODin1 and ODin2 Estimators

For the ODin1 estimator $\tilde{\mathbf{G}}_{w_0,1}$, $h \propto N^{\frac{1}{2d}}$ and the parametric convergence rate is guaranteed when $s \geq d$. This can be achieved with $L \geq d$ parameters and applies to any functional g in (1) that is infinitely differentiable.

In contrast, for the ODin2 estimator $\tilde{\mathbf{G}}_{w_0,2}$, $h \propto N^{\frac{1}{d+1}}$ if $g(x, y)$ has mixed derivatives of the form of $x^\alpha y^\beta$ and the parametric convergence rate is guaranteed when $s \geq \frac{d+1}{2}$. Thus the parametric rate can be achieved with $\tilde{\mathbf{G}}_{w_0,2}$ under less strict assumptions on the smoothness of the densities than those required for $\tilde{\mathbf{G}}_{w_0,1}$. For large d the condition $s \geq (d + 1)/2$ is just slightly stronger than the condition $s \geq d/2$ required by the more complex estimators that achieve the parametric rate proposed in [45].

These rate improvements come at a cost in the number of parameters L required to implement the weighted ensemble estimator. If $s \geq \frac{d+1}{2}$ then the size of J for ODin2 is on the order of $d^2/8$. This may lead to increased variance of the ensemble estimator as indicated by (6). Also, so far $\tilde{\mathbf{G}}_{w_0,2}$ can only be applied to functionals $g(x, y)$ with mixed derivatives of the form of $x^\alpha y^\beta$. Future work is required to extend this estimator to other functionals of interest.

E. Central Limit Theorem

The following theorem shows that the appropriately normalized ensemble estimator $\tilde{\mathbf{G}}_w$ converges in distribution to a normal random variable. This enables us to perform hypothesis testing on the divergence functional. The proof is based on the Efron-Stein inequality and an application of Slutsky's Theorem (Appendix D).

Theorem 5: Assume that the functional g is Lipschitz in both arguments with Lipschitz constant C_g . Further assume that $h = o(1)$, $N \rightarrow \infty$, and $Nh^d \rightarrow \infty$. Then for fixed L , the asymptotic distribution of the weighted ensemble estimator $\tilde{\mathbf{G}}_w$ is

$$Pr\left(\left(\tilde{\mathbf{G}}_w - \mathbb{E}[\tilde{\mathbf{G}}_w]\right) / \sqrt{\mathbb{V}[\tilde{\mathbf{G}}_w]} \leq t\right) \rightarrow Pr(\mathbf{S} \leq t),$$

where \mathbf{S} is a standard normal random variable.

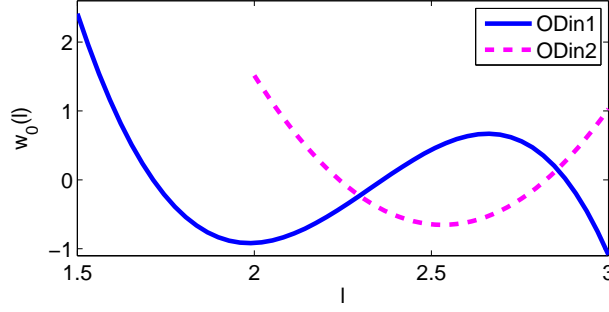


Figure 2. Examples of the optimal weights for $g(x, y) = \left(\frac{x}{y}\right)^\alpha$, $d = 4$, $N = 3100$, $L = 50$, and l is uniformly spaced between 1.5 (ODin1) or 2 (ODin2) and 3. The lowest values of l are given the highest weight. Thus the minimum value of bandwidth parameters \mathcal{L} should be sufficiently large to render an adequate estimate.

IV. NUMERICAL VALIDATION

A. Tuning Parameter Selection

The optimization problem in (7) has parameters η , L , and \mathcal{L} . The parameter η provides an upper bound on the norm of the weight vector, which gives an upper bound on the constant in the variance of the ensemble estimator. If all the constants in (3) or (4) and an exact expression for the variance of the ensemble estimator were known, then η could be chosen to minimize the MSE. Since the constants are unknown, by applying (7), the resulting MSE of the ensemble estimator is $O(\epsilon^2/N) + O(L\eta^2/N)$, where each term in the sum comes from the bias and variance, respectively. Since there is a tradeoff between η and ϵ , in principle setting $\eta = \epsilon/\sqrt{L}$ would minimize these terms. In practice, we find that the variance of the ensemble estimator is less than the upper bound of $L\eta^2/N$ and setting $\eta = \epsilon/\sqrt{L}$ is therefore overly restrictive. Setting $\eta = \epsilon$ instead works well in practice.

For fixed L , the set of kernel widths \mathcal{L} can in theory be chosen by minimizing ϵ in (7) over \mathcal{L} in addition to w . However, this results in a nonconvex optimization problem since w does not lie in the non-negative orthant. A parameter search may not be practical as ϵ generally decreases as the size and spread of \mathcal{L} increases. This decrease in ϵ does not always correspond to a decrease in MSE as high and low values of $h(l)$ can lead to inaccurate density estimates. Denote the value of the minimum value of l so that $\tilde{f}_{i,h(l_{min})}(\mathbf{X}_j) > 0 \forall i = 1, 2$ as l_{min} and the diameter of the support \mathcal{S} as D . To ensure the density estimates are bounded away from zero, we require that $\min(\mathcal{L}) \geq l_{min}$. The weights in w_0 are generally largest for the smallest values of \mathcal{L} (see Fig. 2) so $\min(\mathcal{L})$ should also be sufficiently larger than l_{min} to render an adequate estimate. Similarly, $\max(\mathcal{L})$ should be sufficiently smaller than D as high bandwidth values lead to high bias. The remaining \mathcal{L} values are chosen to be equally spaced between $\min(\mathcal{L})$ and $\max(\mathcal{L})$.

As L increases, the similarity of bandwidth values $h(l)$ and basis functions $\psi_{i,d}(l)$ increases, resulting in a negligible decrease in the bias. Hence L should be chosen large enough to decrease the bias but small enough so that the $h(l)$ values are sufficiently distinct (typically $30 \leq L \leq 60$).

B. Convergence Rates Validation: Rényi- α Divergence

To validate our theory, we estimated the Rényi- α divergence integral between two truncated multivariate Gaussian distributions with varying dimension and sample sizes. The densities have means $\bar{\mu}_1 = 0.7 * \bar{1}_d$, $\bar{\mu}_2 = 0.3 * \bar{1}_d$ and covariance matrices $0.1 * I_d$ where $\bar{1}_d$ is a d -dimensional vector of ones, and I_d is a $d \times d$ identity matrix. We used $\alpha = 0.5$ and restricted the Gaussians to the unit cube.

The left plots in Fig. 3 show the MSE (200 trials) of the standard plug-in estimator implemented with a uniform kernel, the two proposed optimally weighted estimators ODin1 and ODin2, and a linear combination of ODin1 and ODin2, $\tilde{\mathbf{G}}_\rho = (1 - \rho)\tilde{\mathbf{G}}_{w_0,1} + \rho\tilde{\mathbf{G}}_{w_0,2}$, for various dimensions and sample sizes. The set of kernel widths \mathcal{L} , L , and ρ are tuned to minimize the MSE. The bandwidth used for the standard plug-in estimator was selected from the set \mathcal{L} that resulted from the ODin2 optimization; specifically the member of the set that empirically minimized the MSE of the plug-in estimator. Note that for $d = 4$, the standard plug-in estimator performs comparably with the optimally weighted estimators. However, for $d = 7, 10$, the plug-in estimator performs considerably worse. This reflects the strength of ensemble estimators: the weighted sum of a set of poor estimators can result in a very good estimator. Note also that for most cases, the ensemble estimators' MSE rates match the theoretical rate based on the estimated log-log slope given in Table I.

ODin1 tends to do better than ODin2 when the dimension is lower ($d = 4$) while the opposite occurs for the higher dimensions. Further evidence for this is given in the right figures in Fig. 3 that show the corresponding average estimates with standard error bars compared to the true values. ODin1 has smaller variance than ODin2 when $d = 4$ and slightly larger

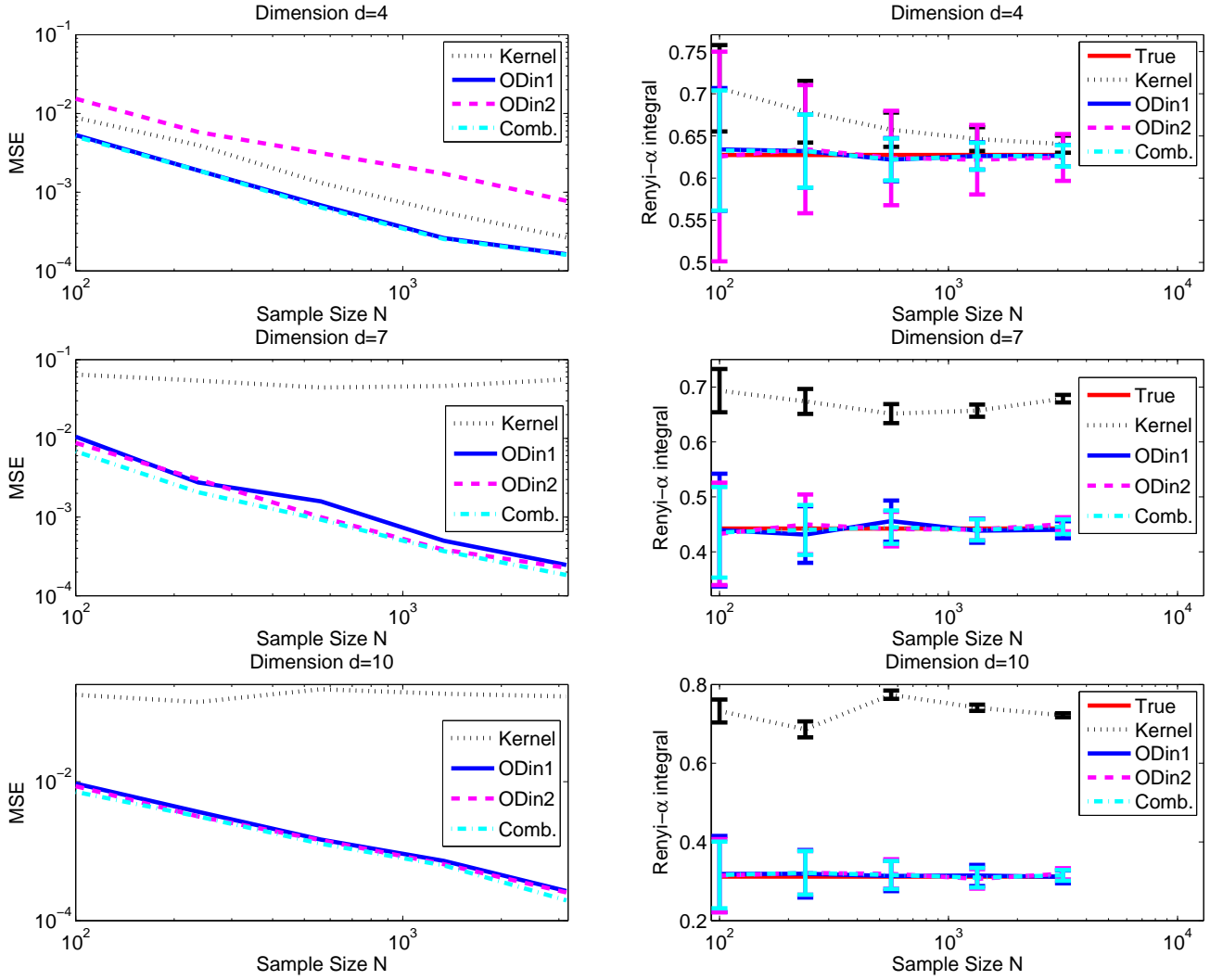


Figure 3. (Left) Log-log plot of MSE of the uniform kernel plug-in (“Kernel”), the two proposed optimally weighted estimators (ODin1 and ODin2), and the optimal linear combination of ODin1 and ODin2 for various dimensions and sample sizes. (Right) Plot of the average value of the same estimators with standard error bars compared to the true values being estimated. The proposed weighted ensemble estimators generally match the theoretical rate (see Table I) and perform much better than the plug-in estimator for high dimensions.

Table I
NEGATIVE LOG-LOG SLOPE OF THE MSE AS A FUNCTION OF SAMPLE SIZE FOR VARIOUS DIMENSIONS AND ESTIMATORS

Estimator	$d = 4$	$d = 7$	$d = 10$
ODin1	1.04	1.07	1.01
ODin2	0.83	1.08	1.00
Comb.	1.03	1.04	1.02

variance when $d = 10$. This seems to account for the differences in MSE between ODin1 and ODin2. The values for the weight ρ are given in Table II which indicate a preference for ODin1 when $d = 4$ and a preference for ODin2 for higher dimensions. Paired t-tests on the MSE (125 trials) of the two methods indicate that the MSE differences are statistically significant (see Table III).

C. Tuning Parameter Robustness

The results in Section IV-B were obtained by selecting the tuning parameters \mathcal{L} and L for each pair of dimension and samples to minimize the MSE. Here we demonstrate the robustness of the estimators to variations in the tuning parameters.

In all experiments, we estimated the Rényi- α divergence integral between the same distributions described in Section IV-B (truncated Gaussians with same covariance and different mean) and chose $L = 50$. In the first set of experiments, we set $\eta = \epsilon$

Table II
VALUES OF THE WEIGHT ρ FOR THE ESTIMATOR $\tilde{\mathbf{G}}_\rho = (1 - \rho)\tilde{\mathbf{G}}_{w_0,1} + \rho\tilde{\mathbf{G}}_{w_0,2}$ THAT MINIMIZE MSE

Dim.	$N = 100$	$N = 240$	$N = 560$	$N = 1330$	$N = 3200$
$d = 4$	0.15	0	0.1	0.05	0.05
$d = 7$	0.6	0.45	0.75	0.75	0.55
$d = 10$	0.55	1	0.5	0.65	0.5

Table III
 p -VALUES OF PAIRED T-TESTS OF ODin1 VS. ODin2 MSE ($N = 1300$).

Dim.	ODin1 > ODin2	ODin1 < ODin2	ODin1 = ODin2
4	1	0	1.8×10^{-58}
7	8.7×10^{-52}	1	1.8×10^{-51}
10	0	1	1.0×10^{-52}

and chose the set of kernel bandwidths \mathcal{L} to be linearly spaced between $\min(\mathcal{L})$ and $\max(\mathcal{L})$. Table IV provides the values chosen for $\min(\mathcal{L})$ and $\max(\mathcal{L})$.

Figure 4 shows the results for these experiments when $d = 5$. As the number of samples increase, choosing a larger range of values for \mathcal{L} (Sets 1 and 2) generally gives better performance for both ODin1 and ODin2 in terms of MSE than choosing a smaller range for \mathcal{L} (e.g. Sets 4 and 5). This suggests that for large sample sizes, the estimators will perform well if a reasonably large range for \mathcal{L} is chosen. In contrast, choosing a smaller range of values for \mathcal{L} when the sample size is small results in smaller bias and variance compared to the larger ranges. Thus for small sample sizes, it may be useful to tighten the range of kernel bandwidths.

Comparing the results for ODin1 and ODin2 indicates that ODin2 is more robust to the choice of \mathcal{L} as the difference in MSE under the different settings is smaller for ODin2 than ODin1. This is due primarily to the relatively smaller bias of ODin2 as the variances of the two estimators under each setting are comparable (see the bottom plots in Fig. 4). Similar results hold when the dimension is increased to $d = 7$ (see Fig. 5). For larger sample sizes, a large range for \mathcal{L} gives better results than a smaller range. However, for smaller sample sizes, the larger range for \mathcal{L} does not perform as well as other configurations. Additionally, ODin2 again appears to be more robust to the choice of \mathcal{L} as the difference in MSE at larger sample sizes is smaller for ODin2. Additionally, the Set 5 configuration of ODin1 does not even appear to be converging to the true value yet when $N = 10000$.

For the second set of experiments, we fixed \mathcal{L} to be linearly spaced values between 2 and 3. We then varied the values of η from 0.5 to 10. Figure 6 provides heatmaps of the MSE of the two ensemble estimators under this configuration with $d = 5, 7$. For $d = 5$, choosing $\eta = 0.5$ gives the lowest MSE when $N \geq 10^{3.5}$ for ODin1 and for all sample sizes for ODin2. In fact, when $d = 5$, ODin2 with $\eta = 0.5$ outperforms all other configurations at all sample sizes, including those shown in Fig. 4. Increasing d to 7 changes this somewhat as choosing $\eta = 0.5$ results in the lowest MSE when $N \geq 1000$ for ODin2 and for no sample sizes for ODin1. However, generally lower values of η ($\eta < 2$) result in the lowest MSE for ODin2 when $N < 1000$ and for ODin1 when $N \geq 10^{3.5}$ ($\eta \leq 3$). Both ODin1 and ODin2 are fairly robust to the choice of η when $d = 5$ as the MSE is relatively constant at each sample size for most η values. However, ODin2 has generally lower MSE values for $N \geq 10^{2.5}$ (see Table V). When $d = 7$, ODin2 is more robust than ODin1 for larger samples ($N \geq 1000$).

Overall, based on our experiments, ODin1 has lower MSE on average for smaller sample sizes while ODin2 is generally more robust to the tuning parameters for larger sample sizes. Additionally, choosing a low value for η with ODin2 may result in better performance. Thus unless the sample size is small, we recommend ODin2 over ODin1.

Table IV
VALUES OF $\min(\mathcal{L})$ AND $\max(\mathcal{L})$ FOR DIFFERENT EXPERIMENTS.

Set	ODin1		ODin2	
	$\min(\mathcal{L})$	$\max(\mathcal{L})$	$\min(\mathcal{L})$	$\max(\mathcal{L})$
1	1.5	3	2	3
2	1.75	3	2.25	3
3	2	3	2.5	3
4	2.25	3	2.75	3
5	2.5	3	2.75	3.25

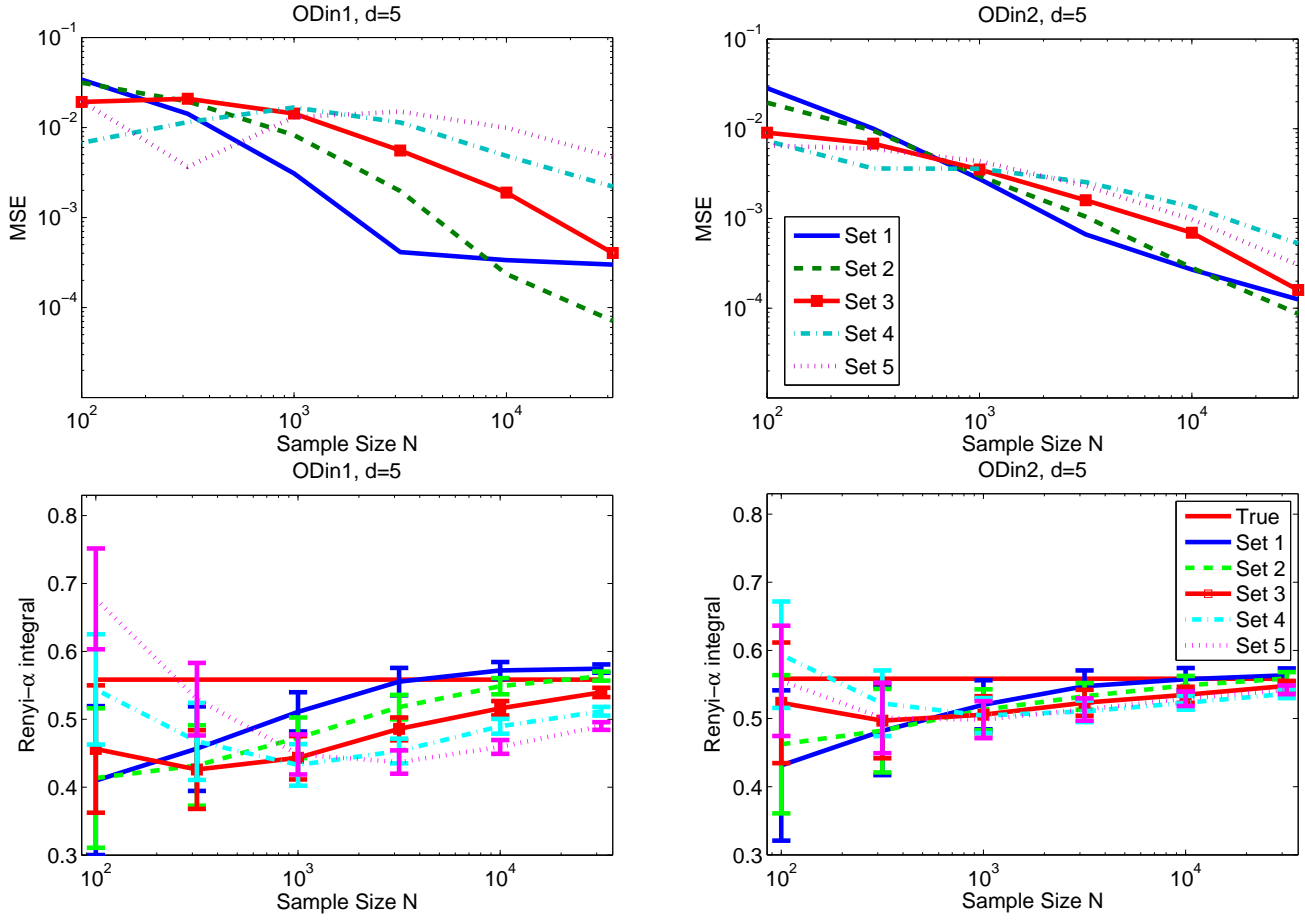


Figure 4. (Top) Log-log plot of MSE of the two proposed optimally weighted estimators (ODin1 and ODin2) as a function of sample size using different values for the range of kernel bandwidths \mathcal{L} (see Table IV) when $d = 5$. (Bottom) Plot of the average value of the same estimators with standard error bars compared to the true value being estimated. For larger sample sizes, a larger range in \mathcal{L} results in smaller MSE (see Sets 1 and 2), while a smaller range in \mathcal{L} is more accurate at smaller sample sizes. ODin2 is generally more robust to the choice of \mathcal{L} .

Table V
AVERAGE MSE OVER ALL VALUES OF η USED IN FIGURE 6 FOR A FIXED SAMPLE SIZE.

Sample Size N	10^2	$10^{2.5}$	10^3	$10^{3.5}$	10^4	$10^{4.5}$
Mean MSE, ODin1, $d = 5$	0.0225	0.0159	0.0118	0.0071	0.0040	0.0022
Mean MSE, ODin2, $d = 5$	0.0358	0.0092	0.0029	0.0011	0.0005	0.0002
Mean MSE, ODin1, $d = 7$	0.0394	0.0233	0.0155	0.0115	0.0091	N/A
Mean MSE, ODin2, $d = 7$	0.0557	0.0292	0.0120	0.0046	0.0018	N/A

D. Central Limit Theorem Validation: KL Divergence

To verify the central limit theorem of both ensemble estimators, we estimated the KL divergence between two truncated Gaussian densities again restricted to the unit cube. We conducted two experiments where 1) the densities are different with means $\bar{\mu}_1 = 0.7 * \mathbf{1}_d$, $\bar{\mu}_2 = 0.3 * \mathbf{1}_d$ and covariance matrices $\sigma_i * I_d$, $\sigma_1 = 0.1$, $\sigma_2 = 0.3$; and where 2) the densities are the same with means $0.3 * \mathbf{1}_d$ and covariance matrices $0.3 * I_d$. For both experiments, we chose $d = 6$ and $N = 1000$.

Figure 7 shows Q-Q plots of the normalized optimally weighted ensemble estimators ODin1 (left) and ODin2 (right) of the KL divergence when the two densities are the same (top) and when they are different (bottom). The linear relationship between the quantiles of the normalized estimators and the standard normal distribution validates Theorem 5 for both estimators under the two cases.

V. CONCLUSION

We derived convergence rates for a kernel density plug-in estimator of divergence functionals. We generalized the theory of optimally weighted ensemble estimation and derived an estimator that achieves the parametric rate when the densities are

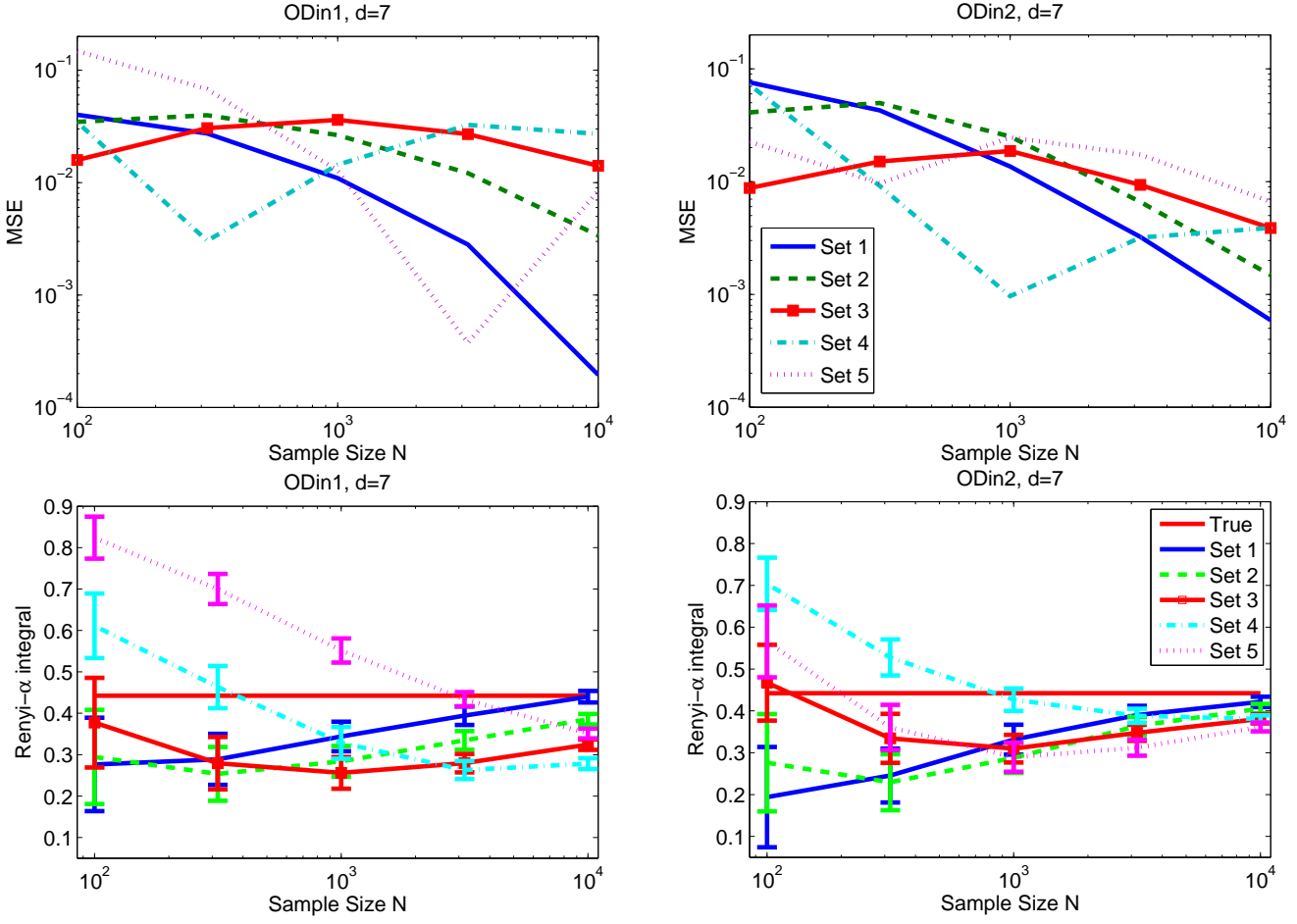


Figure 5. (Top) Log-log plot of MSE of the two proposed optimally weighted estimators (ODin1 and ODin2) as a function of sample size using different values of the parameter \mathcal{L} (see Table IV) when $d = 7$. (Bottom) Plot of the average value of the same estimators with standard error bars compared to the true value being estimated. Again, a larger range for \mathcal{L} at large sample sizes results in smaller MSE.

$(d + 1)/2$ times differentiable. The estimators we derive apply to general bounded density support sets and do not require knowledge of the support which is a distinct advantage over other estimators. We also derived the asymptotic distribution of the estimator, provided some guidelines for tuning parameter selection, and validated the convergence rates for the case of empirical estimation of the Rényi- α divergence. We then performed experiments to examine the estimators' robustness to the choice of tuning parameters and validated the central limit theorem for KL divergence estimation.

Future work includes deriving expressions similar to (4) for more general divergence functionals that are not restricted to the class of functions whose mixed derivatives depend on x, y only through terms of the form $x^\alpha y^\beta$. An important divergence which does not satisfy this condition is the Henze-Penrose divergence [6] which can be used to bound the Bayes error. Further future work will focus on extending this work on distributional functional estimation to k -nn based estimators where knowledge of the support is again not required. This will improve the computational burden as k -nn estimators require fewer computations than standard KDEs.

APPENDIX A PROOF OF THEOREM 1

Consider a uniform rectangular kernel $K(x)$ that satisfies $K(x) = 1$ for all x such that $\|x\|_1 \leq 1/2$. Also consider the family of probability densities f with rectangular support $\mathcal{S} = [-1, 1]^d$. We will prove Theorem 1 which is that that \mathcal{S} satisfies the following smoothness condition (A.5): for any polynomial $p_x(u) : \mathbb{R}^d \rightarrow \mathbb{R}$ of order $q \leq r = \lfloor s \rfloor$ with coefficients that are $r - q$ times differentiable wrt x ,

$$\int_{x \in \mathcal{S}} \left(\int_{u: \|u\|_1 \leq \frac{1}{2}, x+uh \notin \mathcal{S}} p_x(u) du \right)^t dx = v_t(h), \quad (9)$$

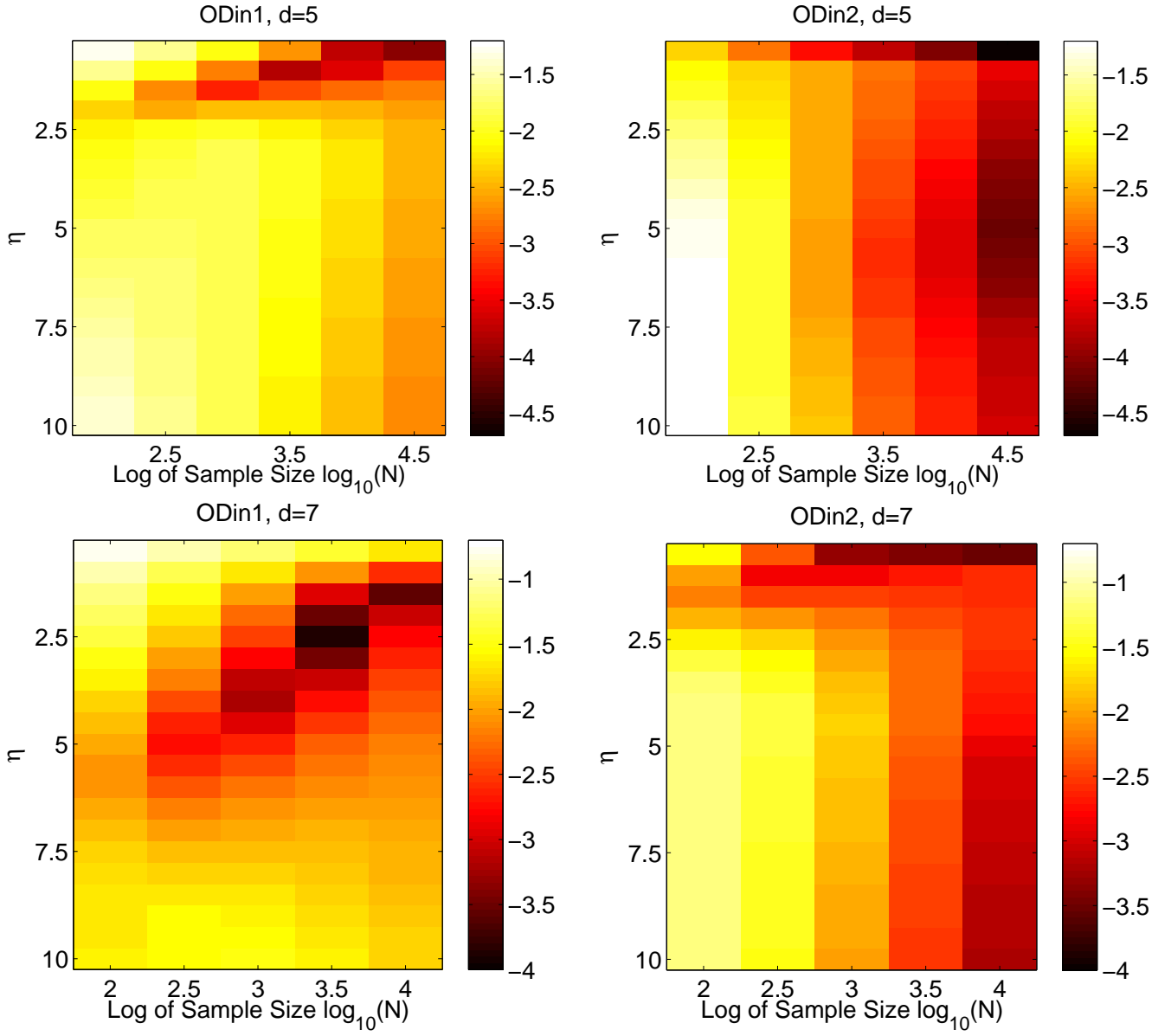


Figure 6. Heatmaps of the ensemble estimators' MSE (\log_{10} scale) as a function of sample size and the tuning parameter η . Lower values of η tend to give the smallest MSE, especially for ODin2. Both estimators are fairly robust to the choice of η as the MSE is relatively constant at each sample size for most η values.

where $v_t(h)$ has the expansion

$$v_t(h) = \sum_{i=1}^{r-q} e_{i,q,t} h^i + o(h^{r-q}).$$

Note that the inner integral forces the x 's under consideration to be boundary points via the constraint $x + uh \notin \mathcal{S}$.

A. Single Coordinate Boundary Point

We begin by focusing on points x that are boundary points by virtue of a single coordinate x_i such that $x_i + u_i h \notin \mathcal{S}$. Without loss of generality, assume that $x_i + u_i h > 1$. The inner integral in (9) can then be evaluated first wrt all coordinates other than i . Since all of these coordinates lie within the support, the inner integral over these coordinates will amount to integration of the polynomial $p_x(u)$ over a symmetric $d-1$ dimensional rectangular region $|u_j| \leq \frac{1}{2}$ for all $j \neq i$. This yields a function $\sum_{m=1}^q \tilde{p}_m(x) u_i^m$ where the coefficients $\tilde{p}_m(x)$ are each $r-q$ times differentiable wrt x .

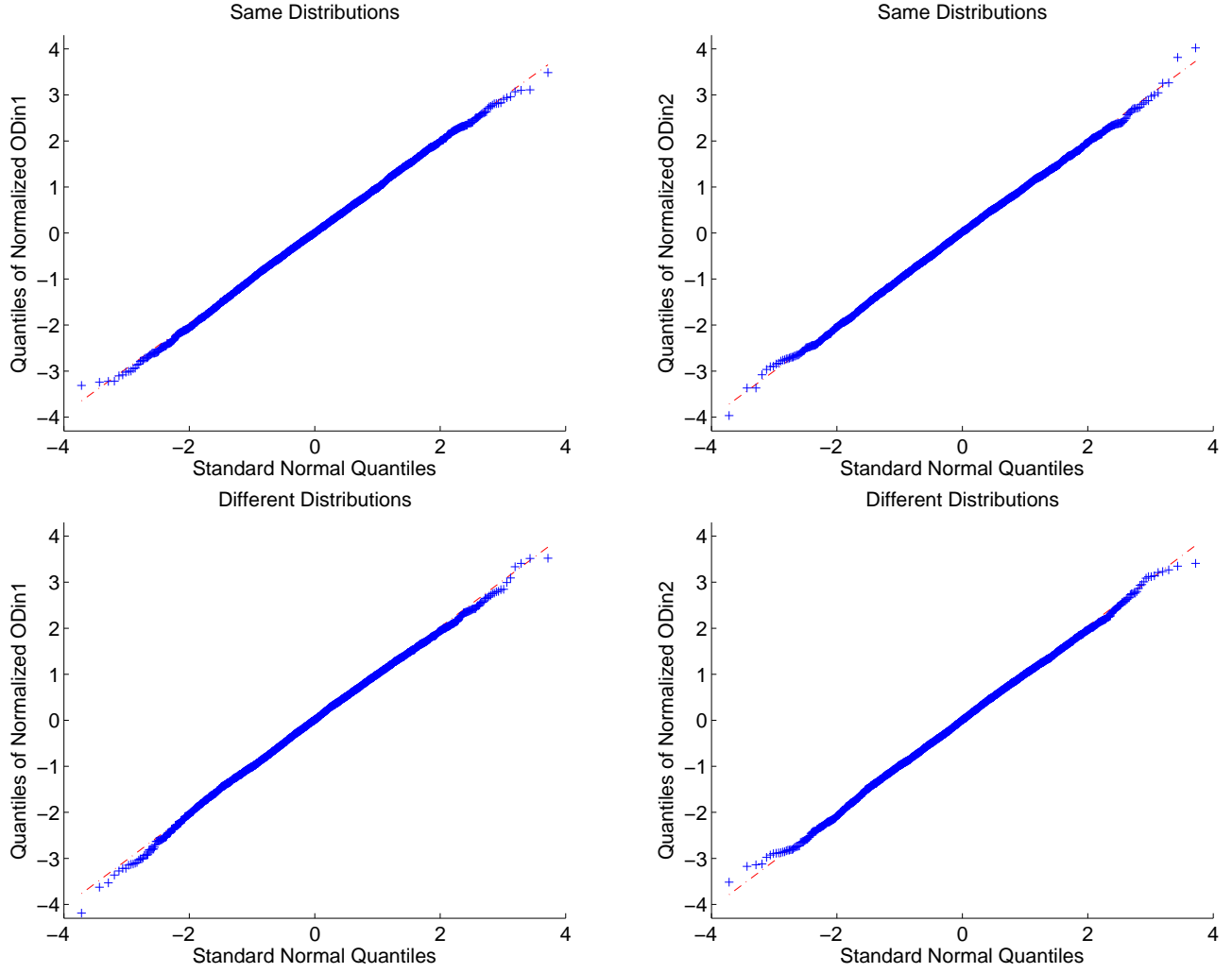


Figure 7. Q-Q plots comparing quantiles from the normalized weighted ensemble estimators ODin1 (left) and ODin2 (right) of the KL divergence (vertical axis) to the quantiles from the standard normal distribution (horizontal axis) when the two distributions are the same (top) and when they are different (bottom). The red line shows a reference line passing through the first and third quantiles. The linearity of the plot points validates the central limit theorem (Theorem 5) for all four cases.

With respect to the u_i coordinate, the inner integral will have limits from $\frac{1-x_i}{h}$ to $\frac{1}{2}$ for some $1 > x_i > 1 - \frac{h}{2}$. Consider the $\tilde{p}_q(x)u_i^q$ monomial term. The inner integral wrt this term yields

$$\sum_{m=1}^q \tilde{p}_m(x) \int_{\frac{1-x_i}{h}}^{\frac{1}{2}} u_i^m du_i = \sum_{m=1}^q \tilde{p}_m(x) \frac{1}{m+1} \left(\frac{1}{2^{m+1}} - \left(\frac{1-x_i}{h} \right)^{m+1} \right). \quad (10)$$

Raising the right hand side of (10) to the power of t results in an expression of the form

$$\sum_{j=0}^{qt} \check{p}_j(x) \left(\frac{1-x_i}{h} \right)^j, \quad (11)$$

where the coefficients $\check{p}_j(x)$ are $r - q$ times differentiable wrt x . Integrating (11) over all the coordinates in x other than x_i results in an expression of the form

$$\sum_{j=0}^{qt} \bar{p}_j(x_i) \left(\frac{1-x_i}{h} \right)^j, \quad (12)$$

where again the coefficients $\bar{p}_j(x_i)$ are $r - q$ times differentiable wrt x_i . Note that since the other coordinates of x other than x_i are far away from the boundary, the coefficients $\bar{p}_j(x_i)$ are independent of h . To evaluate the integral of (12), consider the $r - q$ term Taylor series expansion of $\bar{p}_j(x_i)$ around $x_i = 1$. This will yield terms of the form

$$\begin{aligned} \int_{1-h/2}^1 \frac{(1-x_i)^{j+k}}{h^k} dx_i &= - \frac{(1-x_i)^{j+k+1}}{h^k(j+k+1)} \Big|_{x_i=1-h/2}^{x_i=1} \\ &= \frac{h^{j+1}}{(j+k+1)2^{j+k+1}}, \end{aligned}$$

for $0 \leq j \leq r - q$, and $0 \leq k \leq qt$. Combining terms results in the expansion $v_t(h) = \sum_{i=1}^{r-q} e_{i,q,t} h^i + o(h^{r-q})$.

B. Multiple Coordinate Boundary Point

The case where multiple coordinates of the point x are near the boundary is a straightforward extension of the single boundary point case so we only sketch the main ideas here. As an example, consider the case where 2 of the coordinates are near the boundary. Assume for notational ease that they are x_1 and x_2 and that $x_1 + u_1 h > 1$ and $x_2 + u_2 h > 1$. The inner integral in (9) can again be evaluated first wrt all coordinates other than 1 and 2. This yields a function $\sum_{m,j=1}^q \tilde{p}_{m,j}(x) u_1^m u_2^j$ where the coefficients $\tilde{p}_{m,j}(x)$ are each $r - q$ times differentiable wrt x . Integrating this wrt x_1 and x_2 and then raising the result to the power of t yields a double sum similar to (11). Integrating this over all the coordinates in x other than x_1 and x_2 gives a double sum similar to (12). Then a Taylor series expansion of the coefficients and integration over x_1 and x_2 yields the result.

APPENDIX B PROOF OF THEOREM 2

In this appendix, we prove the bias results in Thm. 2. The bias of the base kernel density plug-in estimator $\tilde{\mathbf{G}}_{h_1, h_2}$ can be expressed as

$$\begin{aligned} \mathbb{B} [\tilde{\mathbf{G}}_{h_1, h_2}] &= \mathbb{E} \left[g(\tilde{\mathbf{f}}_{1, h_1}(\mathbf{Z}), \tilde{\mathbf{f}}_{2, h_2}(\mathbf{Z})) - g(f_1(\mathbf{Z}), f_2(\mathbf{Z})) \right] \\ &= \mathbb{E} \left[g(\tilde{\mathbf{f}}_{1, h_1}(\mathbf{Z}), \tilde{\mathbf{f}}_{2, h_2}(\mathbf{Z})) - g(\mathbb{E}_{\mathbf{Z}} \tilde{\mathbf{f}}_{1, h_1}(\mathbf{Z}), \mathbb{E}_{\mathbf{Z}} \tilde{\mathbf{f}}_{2, h_2}(\mathbf{Z})) \right] \\ &\quad + \mathbb{E} \left[g(\mathbb{E}_{\mathbf{Z}} \tilde{\mathbf{f}}_{1, h_1}(\mathbf{Z}), \mathbb{E}_{\mathbf{Z}} \tilde{\mathbf{f}}_{2, h_2}(\mathbf{Z})) - g(f_1(\mathbf{Z}), f_2(\mathbf{Z})) \right], \end{aligned} \quad (13)$$

where \mathbf{Z} is drawn from f_2 . The first term is the “variance” term while the second is the “bias” term. We bound these terms using Taylor series expansions under the assumption that g is infinitely differentiable. The Taylor series expansion of the variance term in (13) will depend on variance-like terms of the KDEs while the Taylor series expansion of the bias term in (13) will depend on the bias of the KDEs.

The Taylor series expansion of $g(\mathbb{E}_{\mathbf{Z}} \tilde{\mathbf{f}}_{1, h_1}(\mathbf{Z}), \mathbb{E}_{\mathbf{Z}} \tilde{\mathbf{f}}_{2, h_2}(\mathbf{Z}))$ around $f_1(\mathbf{Z})$ and $f_2(\mathbf{Z})$ is

$$g(\mathbb{E}_{\mathbf{Z}} \tilde{\mathbf{f}}_{1, h_1}(\mathbf{Z}), \mathbb{E}_{\mathbf{Z}} \tilde{\mathbf{f}}_{2, h_2}(\mathbf{Z})) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \left(\frac{\partial^{i+j} g(x, y)}{\partial x^i \partial y^j} \Big|_{\substack{x=f_1(\mathbf{Z}) \\ y=f_2(\mathbf{Z})}} \right) \frac{\mathbb{B}_{\mathbf{Z}}^i [\tilde{\mathbf{f}}_{1, h_1}(\mathbf{Z})] \mathbb{B}_{\mathbf{Z}}^j [\tilde{\mathbf{f}}_{2, h_2}(\mathbf{Z})]}{i! j!} \quad (14)$$

where $\mathbb{B}_{\mathbf{Z}}^j [\tilde{\mathbf{f}}_{i, h_i}(\mathbf{Z})] = (\mathbb{E}_{\mathbf{Z}} \tilde{\mathbf{f}}_{i, h_i}(\mathbf{Z}) - f_i(\mathbf{Z}))^j$ is the bias of $\tilde{\mathbf{f}}_{i, h_i}$ at the point \mathbf{Z} raised to the power of j . This expansion can be used to control the second term (the bias term) in (13). To accomplish this, we require an expression for $\mathbb{E}_{\mathbf{Z}} \tilde{\mathbf{f}}_{i, h_i}(\mathbf{Z}) - f_i(\mathbf{Z}) = \mathbb{B}_{\mathbf{Z}} [\tilde{\mathbf{f}}_{i, h_i}(\mathbf{Z})]$.

To obtain an expression for $\mathbb{B}_{\mathbf{Z}} [\tilde{\mathbf{f}}_{i, h_i}(\mathbf{Z})]$, we consider separately the cases when \mathbf{Z} is in the interior of the support \mathcal{S} or when \mathbf{Z} is near the boundary of the support. A point $X \in \mathcal{S}$ is defined to be in the interior of \mathcal{S} if for all $Y \notin \mathcal{S}$, $K\left(\frac{X-Y}{h_i}\right) = 0$. A point $X \in \mathcal{S}$ is near the boundary of the support if it is not in the interior. Denote the region in the interior and near the boundary wrt h_i as \mathcal{S}_{I_i} and \mathcal{S}_{B_i} , respectively. We will need the following.

Lemma 1: Let \mathbf{Z} be a realization of the density f_2 independent of $\tilde{\mathbf{f}}_{i, h_i}$ for $i = 1, 2$. Assume that the densities f_1 and f_2 belong to $\Sigma(s, L)$. Then for $\mathbf{Z} \in \mathcal{S}_{I_i}$,

$$\mathbb{B}_{\mathbf{Z}} [\tilde{\mathbf{f}}_{i, h_i}(\mathbf{Z})] = f_i(\mathbf{Z}) + \sum_{j=\nu/2}^{\lfloor s/2 \rfloor} c_{i,j}(\mathbf{Z}) h_i^{2j} + O(h_i^s). \quad (15)$$

Proof: Obtaining the lower order terms in (15) is a common result in kernel density estimation. However, since we also require the higher order terms, we present the proof here. Additionally, some of the results in this proof will be useful later. From the linearity of the KDE, we have that if \mathbf{X} is drawn from f_i and is independent of \mathbf{Z} , then

$$\begin{aligned}\mathbb{E}_{\mathbf{Z}} \tilde{\mathbf{f}}_{i,h_i}(\mathbf{Z}) &= \mathbb{E}_{\mathbf{Z}} \left[\frac{1}{h_i^d} K \left(\frac{\mathbf{X} - \mathbf{Z}}{h_i} \right) \right] \\ &= \int \frac{1}{h_i^d} K \left(\frac{x - \mathbf{Z}}{h_i} \right) f_i(x) dx \\ &= \int K(t) f_i(th_i + \mathbf{Z}) dt,\end{aligned}\tag{16}$$

where the last step follows from the substitution $t = \frac{x - \mathbf{Z}}{h_i}$. Since the density f_i belongs to $\Sigma(s, K)$, using multi-index notation we can expand it as

$$f_i(th_i + \mathbf{Z}) = f_i(\mathbf{Z}) + \sum_{0 < |\alpha| \leq \lfloor s \rfloor} \frac{D^\alpha f_i(\mathbf{Z})}{\alpha!} (th_i)^\alpha + O(\|th_i\|^s),\tag{17}$$

where $\alpha! = \alpha_1! \alpha_2! \dots \alpha_d!$ and $Y^\alpha = Y_1^{\alpha_1} Y_2^{\alpha_2} \dots Y_d^{\alpha_d}$. Combining (16) and (17) gives

$$\begin{aligned}\mathbb{E}_{\mathbf{Z}} \tilde{\mathbf{f}}_{i,h_i}(\mathbf{Z}) &= f_i(\mathbf{Z}) + \sum_{0 < |\alpha| \leq \lfloor s \rfloor} \frac{D^\alpha f_i(\mathbf{Z})}{\alpha!} h_i^{|\alpha|} \int t^\alpha K(t) dt + O(h_i^s) \\ &= f_i(\mathbf{Z}) + \sum_{j=\nu/2}^{\lfloor s/2 \rfloor} c_{i,j}(\mathbf{Z}) h_i^{2j} + O(h_i^s),\end{aligned}$$

where the last step follows from the fact that K is symmetric and of order ν . ■

To obtain a similar result for the case when \mathbf{Z} is near the boundary of \mathcal{S} , we use assumption $\mathcal{A}.5$.

Lemma 2: Let $\gamma(x, y)$ be an arbitrary function satisfying $\sup_{x,y} |\gamma(x, y)| < \infty$. Let \mathcal{S} satisfy the boundary smoothness conditions of Assumption $\mathcal{A}.5$. Assume that the densities f_1 and f_2 belong to $\Sigma(s, L)$ and let \mathbf{Z} be a realization of the density f_2 independent of $\tilde{\mathbf{f}}_{i,h_i}$ for $i = 1, 2$. Let $h' = \min(h_1, h_2)$. Then

$$\mathbb{E} \left[1_{\{\mathbf{Z} \in \mathcal{S}_{B_i}\}} \gamma(f_1(\mathbf{Z}), f_2(\mathbf{Z})) \mathbb{B}_{\mathbf{Z}}^t \left[\tilde{\mathbf{f}}_{i,h_i}(\mathbf{Z}) \right] \right] = \sum_{j=1}^r c_{4,i,j,t} h_i^j + o(h_i^r)\tag{18}$$

$$\mathbb{E} \left[1_{\{\mathbf{Z} \in \mathcal{S}_{B_1} \cap \mathcal{S}_{B_2}\}} \gamma(f_1(\mathbf{Z}), f_2(\mathbf{Z})) \mathbb{B}_{\mathbf{Z}}^t \left[\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}) \right] \mathbb{B}_{\mathbf{Z}}^q \left[\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z}) \right] \right] = \sum_{j=0}^{r-1} \sum_{i=0}^{r-1} c_{4,j,i,q,t} h_1^j h_2^i h' + o\left((h')^r\right)\tag{19}$$

Proof: For fixed X near the boundary of \mathcal{S} , we have

$$\begin{aligned}\mathbb{E} \left[\tilde{\mathbf{f}}_{i,h_i}(X) \right] - f_i(X) &= \frac{1}{h_i^d} \int_{Y:Y \in \mathcal{S}} K \left(\frac{X - Y}{h_i} \right) f_i(Y) dY - f_i(X) \\ &= \left[\frac{1}{h_i^d} \int_{Y:K\left(\frac{X-Y}{h_i}\right) > 0} K \left(\frac{X - Y}{h_i} \right) f_i(Y) dY - f_i(X) \right] \\ &\quad - \left[\frac{1}{h_i^d} \int_{Y:Y \notin \mathcal{S}} K \left(\frac{X - Y}{h_i} \right) f_i(Y) dY \right] \\ &= T_{1,i}(X) - T_{2,i}(X).\end{aligned}$$

Note that in $T_{1,i}(X)$, we are extending the integral beyond the support of the density f_i . However, by using the same Taylor series expansion method as in the proof of Lemma 1, we always evaluate f_i and its derivatives at the point X which is within the support of f_i . Thus it does not matter how we define an extension of f_i since the Taylor series will remain the same. Thus $T_{1,i}(X)$ results in an identical expression to that obtained from (15).

For the $T_{2,i}(X)$ term, we expand it as follows using multi-index notation as

$$\begin{aligned}T_{2,i}(X) &= \frac{1}{h_i^d} \int_{Y:Y \notin \mathcal{S}} K \left(\frac{X - Y}{h_i} \right) f_i(Y) dY \\ &= \int_{u:h_i u + X \notin \mathcal{S}, K(u) > 0} K(u) f_i(X + h_i u) du \\ &= \sum_{|\alpha| \leq r} \frac{h_i^{|\alpha|}}{\alpha!} \int_{u:h_i u + X \notin \mathcal{S}, K(u) > 0} K(u) D^\alpha f_i(X) u^\alpha du + o(h_i^r).\end{aligned}$$

Recognizing that the $|\alpha|$ th derivative of f_i is $r - |\alpha|$ times differentiable, we can apply assumption $\mathcal{A}.5$ to obtain the expectation of $T_{2,i}(X)$ wrt X :

$$\begin{aligned}
\mathbb{E}[T_{2,i}(\mathbf{X})] &= \frac{1}{h_i^d} \int_X \int_{Y: Y \notin \mathcal{S}} K\left(\frac{X-Y}{h_i}\right) f_i(Y) dY f_2(X) dx \\
&= \sum_{|\alpha| \leq r} \frac{h_i^{|\alpha|}}{\alpha!} \int_X \int_{u: h_i u + X \notin \mathcal{S}, K(u) > 0} K(u) D^\alpha f_i(X) u^\alpha du f_2(X) dX + o(h_i^r) \\
&= \sum_{|\alpha| \leq r} \frac{h_i^{|\alpha|}}{\alpha!} \left[\sum_{1 \leq |\beta| \leq r - |\alpha|} e_{\beta, r - |\alpha|} h_i^{|\beta|} + o(h_i^{r - |\alpha|}) \right] + o(h_i^r) \\
&= \sum_{j=1}^r e_j h_i^j + o(h_i^r).
\end{aligned}$$

Similarly, we find that

$$\begin{aligned}
\mathbb{E}[(T_{2,i}(\mathbf{X}))^t] &= \frac{1}{h_i^{dt}} \int_X \left(\int_{Y: Y \notin \mathcal{S}} K\left(\frac{X-Y}{h_i}\right) f_i(Y) dY \right)^t f_2(X) dx \\
&= \int_X \left(\sum_{|\alpha| \leq r} \frac{h_i^{|\alpha|}}{\alpha!} \int_{u: h_i u + X \notin \mathcal{S}, K(u) > 0} K(u) D^\alpha f_i(X) u^\alpha du \right)^t f_2(X) dX \\
&= \sum_{j=1}^r e_{j,t} h_i^j + o(h_i^r).
\end{aligned}$$

Combining these results gives

$$\begin{aligned}
\mathbb{E} \left[1_{\{\mathbf{Z} \in \mathcal{S}_B\}} \gamma(f_1(\mathbf{Z}), f_2(\mathbf{Z})) \left(\mathbb{E}_{\mathbf{Z}} [\tilde{\mathbf{f}}_{i,h_i}(\mathbf{Z})] - f_i(\mathbf{Z}) \right)^t \right] &= \mathbb{E} \left[\gamma(f_1(\mathbf{Z}), f_2(\mathbf{Z})) (T_{1,i}(\mathbf{Z}) - T_{2,i}(\mathbf{Z}))^t \right] \\
&= \mathbb{E} \left[\gamma(f_1(\mathbf{Z}), f_2(\mathbf{Z})) \sum_{j=0}^t \binom{t}{j} (T_{1,i}(\mathbf{Z}))^j (-T_{2,i}(\mathbf{Z}))^{t-j} \right] \\
&= \sum_{j=1}^r c_{4,i,j,t} h_i^j + o(h_i^r),
\end{aligned}$$

where the constants are functionals of the kernel, γ , and the densities.

The expression in (19) can be proved in a similar manner. ■

Applying Lemmas 1 and 2 to (14) gives

$$\mathbb{E} \left[g \left(\mathbb{E}_{\mathbf{Z}} \tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}), \mathbb{E}_{\mathbf{Z}} \tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z}) \right) - g(f_1(\mathbf{Z}), f_2(\mathbf{Z})) \right] = \sum_{j=1}^r \left(c_{4,1,j} h_1^j + c_{4,2,j} h_2^j \right) + \sum_{j=0}^{r-1} \sum_{i=0}^{r-1} c_{5,i,j} h_1^j h_2^i h' + o(h_1^r + h_2^r). \quad (20)$$

For the variance term (the first term) in (13), the truncated Taylor series expansion of $g(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z}))$ around $\mathbb{E}_{\mathbf{Z}} \tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z})$ and $\mathbb{E}_{\mathbf{Z}} \tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})$ gives

$$g(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})) = \sum_{i=0}^{\lambda} \sum_{j=0}^{\lambda} \left(\frac{\partial^{i+j} g(x, y)}{\partial x^i \partial y^j} \Big|_{\substack{x = \mathbb{E}_{\mathbf{Z}} \tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}) \\ y = \mathbb{E}_{\mathbf{Z}} \tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})}} \right) \frac{\tilde{\mathbf{e}}_{1,h_1}^i(\mathbf{Z}) \tilde{\mathbf{e}}_{2,h_2}^j(\mathbf{Z})}{i! j!} + o(\tilde{\mathbf{e}}_{1,h_1}^\lambda(\mathbf{Z}) + \tilde{\mathbf{e}}_{2,h_2}^\lambda(\mathbf{Z})) \quad (21)$$

where $\tilde{\mathbf{e}}_{i,h_i}(\mathbf{Z}) := \tilde{\mathbf{f}}_{i,h_i}(\mathbf{Z}) - \mathbb{E}_{\mathbf{Z}} \tilde{\mathbf{f}}_{i,h_i}(\mathbf{Z})$. To control the variance term in (13), we thus require expressions for $\mathbb{E}_{\mathbf{Z}} [\tilde{\mathbf{e}}_{i,h_i}^j(\mathbf{Z})]$.

Lemma 3: Let \mathbf{Z} be a realization of the density f_2 that is in the interior of the support and is independent of $\tilde{\mathbf{f}}_{i,h_i}$ for $i = 1, 2$. Let $n(q)$ be the set of integer divisors of q including 1 but excluding q . Then,

$$\mathbb{E}_{\mathbf{Z}} \left[\tilde{\mathbf{e}}_{i,h_i}^q(\mathbf{Z}) \right] = \begin{cases} \sum_{j \in n(q)} \frac{1}{(N_2 h_2^d)^{q-j}} \sum_{m=0}^{\lfloor s/2 \rfloor} c_{6,i,q,j,m}(\mathbf{Z}) h_i^{2m} + O\left(\frac{1}{N_i}\right), & q \geq 2 \\ 0, & q = 1, \end{cases} \quad (22)$$

$$\mathbb{E}_{\mathbf{Z}} \left[\tilde{\mathbf{e}}_{1,h_1}^q(\mathbf{Z}) \tilde{\mathbf{e}}_{2,h_2}^l(\mathbf{Z}) \right] = \begin{cases} \left(\sum_{i \in n(q)} \frac{1}{(N_1 h_1^d)^{q-i}} \sum_{m=0}^{\lfloor s/2 \rfloor} c_{6,1,q,i,m}(\mathbf{Z}) h_1^{2m} \right) \times & q, l \geq 2 \\ \left(\sum_{j \in n(l)} \frac{1}{(N_2 h_2^d)^{l-j}} \sum_{t=0}^{\lfloor s/2 \rfloor} c_{6,2,l,j,t}(\mathbf{Z}) h_2^{2t} \right) + O\left(\frac{1}{N_1} + \frac{1}{N_2}\right), & \\ 0, & q = 1 \text{ or } l = 1 \end{cases} \quad (23)$$

where $c_{6,i,q,j,m}$ is a functional of f_1 and f_2 .

Proof: Define the random variable $\mathbf{V}_i(\mathbf{Z}) = K\left(\frac{\mathbf{X}_i - \mathbf{Z}}{h_2}\right) - \mathbb{E}_{\mathbf{Z}} K\left(\frac{\mathbf{X}_i - \mathbf{Z}}{h_2}\right)$. This gives

$$\begin{aligned} \tilde{\mathbf{e}}_{2,h_2}(\mathbf{Z}) &= \tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z}) - \mathbb{E}_{\mathbf{Z}} \tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z}) \\ &= \frac{1}{N_2 h_2^d} \sum_{i=1}^{N_2} \mathbf{V}_i(\mathbf{Z}). \end{aligned}$$

Clearly, $\mathbb{E}_{\mathbf{Z}} \mathbf{V}_i(\mathbf{Z}) = 0$. From (16), we have for integer $j \geq 1$

$$\begin{aligned} \mathbb{E}_{\mathbf{Z}} \left[K^j \left(\frac{\mathbf{X}_i - \mathbf{Z}}{h_2} \right) \right] &= \int K^j(t) f_2(th_2 + \mathbf{Z}) dt \\ &= h_2^d \sum_{m=0}^{\lfloor s/2 \rfloor} c_{3,2,j,m}(\mathbf{Z}) h_2^{2m}, \end{aligned}$$

where the constants $c_{3,2,j,m}$ depend on the density f_2 , its derivatives, and the moments of the kernel K^j . Note that since K is symmetric, the odd moments of K^j are zero for \mathbf{Z} in the interior of the support. However, all even moments may now be nonzero since K^j may now be nonnegative. By the binomial theorem,

$$\begin{aligned} \mathbb{E}_{\mathbf{Z}} \left[\mathbf{V}_i^j(\mathbf{Z}) \right] &= \sum_{k=0}^j \binom{j}{k} \mathbb{E}_{\mathbf{Z}} \left[K^k \left(\frac{\mathbf{X}_i - \mathbf{Z}}{h_2} \right) \right] \mathbb{E}_{\mathbf{Z}} \left[K \left(\frac{\mathbf{X}_i - \mathbf{Z}}{h_2} \right) \right]^{j-k} \\ &= \sum_{k=0}^j \binom{j}{k} h_2^d O\left(h_2^{d(j-k)}\right) \sum_{m=0}^{\lfloor s/2 \rfloor} c_{3,2,k,m}(\mathbf{Z}) h_2^{2m} \\ &= h_2^d \sum_{m=0}^{\lfloor s/2 \rfloor} c_{3,2,j,m}(\mathbf{Z}) h_2^{2m} + O(h_2^{2d}). \end{aligned}$$

We can use these expressions to simplify $\mathbb{E}_{\mathbf{Z}} \left[\tilde{\mathbf{e}}_{2,h_2}^q(\mathbf{Z}) \right]$. As an example, let $q = 2$. Then since the \mathbf{X}_i s are independent,

$$\begin{aligned} \mathbb{E}_{\mathbf{Z}} \left[\tilde{\mathbf{e}}_{2,h_2}^2(\mathbf{Z}) \right] &= \frac{1}{N_2 h_2^{2d}} \mathbb{E}_{\mathbf{Z}} \mathbf{V}_i^2(\mathbf{Z}) \\ &= \frac{1}{N_2 h_2^d} \sum_{m=0}^{\lfloor s/2 \rfloor} c_{3,2,2,m}(\mathbf{Z}) h_2^{2m} + O\left(\frac{1}{N_2}\right). \end{aligned}$$

Similarly, we find that

$$\begin{aligned} \mathbb{E}_{\mathbf{Z}} \left[\tilde{\mathbf{e}}_{2,h_2}^3(\mathbf{Z}) \right] &= \frac{1}{N_2^2 h_2^{3d}} \mathbb{E}_{\mathbf{Z}} \mathbf{V}_i^3(\mathbf{Z}) \\ &= \frac{1}{(N_2 h_2^d)^2} \sum_{m=0}^{\lfloor s/2 \rfloor} c_{3,2,3,m}(\mathbf{Z}) h_2^{2m} + o\left(\frac{1}{N_2}\right). \end{aligned}$$

For $q = 4$, we have

$$\begin{aligned}\mathbb{E}_{\mathbf{Z}} [\tilde{\mathbf{e}}_{2,h_2}^4(\mathbf{Z})] &= \frac{1}{N_2^3 h_2^{4d}} \mathbb{E}_{\mathbf{Z}} \mathbf{V}_i^4(\mathbf{Z}) + \frac{N_2 - 1}{N_2^3 h_2^{4d}} (\mathbb{E}_{\mathbf{Z}} \mathbf{V}_i^2(\mathbf{Z}))^2 \\ &= \frac{1}{(N_2 h_2^d)^3} \sum_{m=0}^{\lfloor s/2 \rfloor} c_{3,2,4,m}(\mathbf{Z}) h_2^{2m} + \frac{1}{(N_2 h_2^d)^2} \sum_{m=0}^{\lfloor s/2 \rfloor} c_{6,2,2,m}(\mathbf{Z}) h_2^{2m} + o\left(\frac{1}{N_2}\right).\end{aligned}$$

The pattern is then for $q \geq 2$,

$$\mathbb{E}_{\mathbf{Z}} [\tilde{\mathbf{e}}_{2,h_2}^q(\mathbf{Z})] = \sum_{i \in n(q)} \frac{1}{(N_2 h_2^d)^{q-i}} \sum_{m=0}^{\lfloor s/2 \rfloor} c_{6,2,q,i,m}(\mathbf{Z}) h_2^{2m} + O\left(\frac{1}{N_2}\right).$$

For any integer q , the largest possible factor is $q/2$. Thus for given q , the smallest possible exponent on the $N_2 h_2^d$ term is $q/2$. This increases as q increases. A similar expression holds for $\mathbb{E}_{\mathbf{Z}} [\tilde{\mathbf{e}}_{1,h_1}^q(\mathbf{Z})]$ except the \mathbf{X}_i s are replaced with \mathbf{Y}_i , f_2 is replaced with f_1 , and N_2 and h_2 are replaced with N_1 and h_1 , respectively, all resulting in different constants. Then since $\tilde{\mathbf{e}}_{1,h_1}(\mathbf{Z})$ and $\tilde{\mathbf{e}}_{2,h_2}(\mathbf{Z})$ are conditionally independent given \mathbf{Z} ,

$$\begin{aligned}\mathbb{E}_{\mathbf{Z}} [\tilde{\mathbf{e}}_{1,h_1}^q(\mathbf{Z}) \tilde{\mathbf{e}}_{2,h_2}^l(\mathbf{Z})] &= \left(\sum_{i \in n(q)} \frac{1}{(N_1 h_1^d)^{q-i}} \sum_{m=0}^{\lfloor s/2 \rfloor} c_{6,1,q,i,m}(\mathbf{Z}) h_1^{2m} \right) \left(\sum_{j \in n(l)} \frac{1}{(N_2 h_2^d)^{l-j}} \sum_{t=0}^{\lfloor s/2 \rfloor} c_{6,2,l,j,t}(\mathbf{Z}) h_2^{2t} \right) \\ &\quad + O\left(\frac{1}{N_1} + \frac{1}{N_2}\right).\end{aligned}$$

■

Applying Lemma 3 to (21) when taking the conditional expectation given \mathbf{Z} in the interior gives an expression of the form

$$\begin{aligned}\sum_{j=1}^{\lambda/2} \sum_{m=0}^{\lfloor s/2 \rfloor} \left(c_{7,1,j,m} \left(\mathbb{E}_{\mathbf{Z}} \tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}), \mathbb{E}_{\mathbf{Z}} \tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z}) \right) \frac{h_1^{2m}}{(N_1 h_1^d)^j} + c_{7,2,j,m} \left(\mathbb{E}_{\mathbf{Z}} \tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z}), \mathbb{E}_{\mathbf{Z}} \tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z}) \right) \frac{h_2^{2m}}{(N_2 h_2^d)^j} \right) \\ + \sum_{j=1}^{\lambda/2} \sum_{m=0}^{\lfloor s/2 \rfloor} \sum_{i=1}^{\lambda/2} \sum_{n=0}^{\lfloor s/2 \rfloor} c_{7,j,i,m,n} \left(\mathbb{E}_{\mathbf{Z}} \tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z}), \mathbb{E}_{\mathbf{Z}} \tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z}) \right) \frac{h_1^{2m} h_2^{2n}}{(N_1 h_1^d)^j (N_2 h_2^d)^i} \\ + O\left(\frac{1}{(N_1 h_1^d)^{\frac{\lambda}{2}}} + \frac{1}{(N_2 h_2^d)^{\frac{\lambda}{2}}}\right).\end{aligned}\quad (24)$$

Note that the functionals $c_{7,i,j,m}$ and $c_{7,j,i,m,n}$ depend on the derivatives of g and $\mathbb{E}_{\mathbf{Z}} \tilde{\mathbf{f}}_{i,h_i}(\mathbf{Z})$ which depends on h_i . To apply ensemble estimation, we need to separate the dependence on h_i from the constants. If we use ODin1, then it is sufficient to note that in the interior of the support, $\mathbb{E}_{\mathbf{Z}} \tilde{\mathbf{f}}_{i,h_i}(\mathbf{Z}) = f_i(\mathbf{Z}) + o(1)$ and therefore $c(\mathbb{E}_{\mathbf{Z}} \tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}), \mathbb{E}_{\mathbf{Z}} \tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})) = c(f_1(\mathbf{Z}), f_2(\mathbf{Z})) + o(1)$ for some functional c . The terms in (24) reduce to

$$c_{7,1,1,0}(f_1(\mathbf{Z}), f_2(\mathbf{Z})) \frac{1}{N_1 h_1^d} + c_{7,2,1,0}(f_1(\mathbf{Z}), f_2(\mathbf{Z})) \frac{1}{N_2 h_2^d} + o\left(\frac{1}{N_1 h_1^d} + \frac{1}{N_2 h_2^d}\right).$$

For ODin2, we need the higher order terms. To separate the dependence on h_i from the constants, we need more information about the functional g and its derivatives. Consider the special case where the functional $g(x, y)$ has derivatives of the form of $x^\alpha y^\beta$ with $\alpha, \beta < 0$. This includes the important cases of the KL divergence and the Renyi divergence. The generalized binomial theorem states that if $\binom{\alpha}{m} := \frac{\alpha(\alpha-1)\dots(\alpha-m+1)}{m!}$ and if q and t are real numbers with $|q| > |t|$, then for any complex number α ,

$$(q + t)^\alpha = \sum_{m=0}^{\infty} \binom{\alpha}{m} q^{\alpha-m} t^m. \quad (25)$$

Since the densities are bounded away from zero, for sufficiently small h_i , we have that $f_i(\mathbf{Z}) > \left| \sum_{j=\nu/2}^{\lfloor s/2 \rfloor} c_{i,j}(\mathbf{Z}) h_i^{2j} + O(h_i^s) \right|$. Applying the generalized binomial theorem and Lemma 1 gives that

$$\left(\mathbb{E}_{\mathbf{Z}} \tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}) \right)^\alpha = \sum_{m=0}^{\infty} \binom{\alpha}{m} f_i^{\alpha-m}(\mathbf{Z}) \left(\sum_{j=\nu/2}^{\lfloor s/2 \rfloor} c_{i,j}(\mathbf{Z}) h_i^{2j} + O(h_i^s) \right)^m.$$

Since m is an integer, the exponents of the h_i terms are also integers. Thus (24) gives in this case

$$\begin{aligned} \mathbb{E}_{\mathbf{Z}} \left[g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z}) \right) - g \left(\mathbb{E}_{\mathbf{Z}} \tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}), \mathbb{E}_{\mathbf{Z}} \tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z}) \right) \right] &= \sum_{j=1}^{\lambda/2} \sum_{m=0}^{\lfloor s/2 \rfloor} \left(c_{8,1,j,m}(\mathbf{Z}) \frac{h_1^{2m}}{(N_1 h_1^d)^j} + c_{8,2,j,m}(\mathbf{Z}) \frac{h_2^{2m}}{(N_2 h_2^d)^j} \right) \\ &+ \sum_{j=1}^{\lambda/2} \sum_{m=0}^{\lfloor s/2 \rfloor} \sum_{i=1}^{\lambda/2} \sum_{n=0}^{\lfloor s/2 \rfloor} c_{8,j,i,m,n}(\mathbf{Z}) \frac{h_1^{2m} h_2^{2n}}{(N_1 h_1^d)^j (N_2 h_2^d)^i} \\ &+ O \left(\frac{1}{(N_1 h_1^d)^{\frac{\lambda}{2}}} + \frac{1}{(N_2 h_2^d)^{\frac{\lambda}{2}}} + h_1^s + h_2^s \right). \end{aligned} \quad (26)$$

As before, the case for \mathbf{Z} close to the boundary of the support is more complicated. However, by using a similar technique to the proof of Lemma 2 for \mathbf{Z} at the boundary and combining with the previous results, we find that for general g ,

$$\mathbb{E} \left[g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z}) \right) - g \left(\mathbb{E}_{\mathbf{Z}} \tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}), \mathbb{E}_{\mathbf{Z}} \tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z}) \right) \right] = c_{9,1} \frac{1}{N_1 h_1^d} + c_{9,2} \frac{1}{N_2 h_2^d} + o \left(\frac{1}{N_1 h_1^d} + \frac{1}{N_2 h_2^d} \right). \quad (27)$$

If $g(x, y)$ has derivatives of the form of $x^\alpha y^\beta$ with $\alpha, \beta < 0$, then we can similarly obtain

$$\begin{aligned} \mathbb{E} \left[g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z}) \right) - g \left(\mathbb{E}_{\mathbf{Z}} \tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}), \mathbb{E}_{\mathbf{Z}} \tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z}) \right) \right] &= \sum_{j=1}^{\lambda/2} \sum_{m=0}^r \left(c_{9,1,j,m} \frac{h_1^m}{(N_1 h_1^d)^j} + c_{9,2,j,m} \frac{h_2^m}{(N_2 h_2^d)^j} \right) \\ &+ \sum_{j=1}^{\lambda/2} \sum_{m=0}^r \sum_{i=1}^{\lambda/2} \sum_{n=0}^r c_{9,j,i,m,n} \frac{h_1^m h_2^n}{(N_1 h_1^d)^j (N_2 h_2^d)^i} \\ &+ O \left(\frac{1}{(N_1 h_1^d)^{\frac{\lambda}{2}}} + \frac{1}{(N_2 h_2^d)^{\frac{\lambda}{2}}} + h_1^s + h_2^s \right). \end{aligned} \quad (28)$$

Combining (20) with either (27) or (28) completes the proof.

APPENDIX C PROOF OF THEOREM 3

To bound the variance of the plug-in estimator $\tilde{\mathbf{G}}_{h_1, h_2}$, we will use the Efron-Stein inequality [58]:

Lemma 4 (Efron-Stein Inequality): Let $\mathbf{X}_1, \dots, \mathbf{X}_n, \mathbf{X}'_1, \dots, \mathbf{X}'_n$ be independent random variables on the space \mathcal{S} . Then if $f : \mathcal{S} \times \dots \times \mathcal{S} \rightarrow \mathbb{R}$, we have that

$$\mathbb{V} [f(\mathbf{X}_1, \dots, \mathbf{X}_n)] \leq \frac{1}{2} \sum_{i=1}^n \mathbb{E} \left[\left(f(\mathbf{X}_1, \dots, \mathbf{X}_n) - f(\mathbf{X}_1, \dots, \mathbf{X}'_i, \dots, \mathbf{X}_n) \right)^2 \right].$$

Suppose we have samples $\{\mathbf{X}_1, \dots, \mathbf{X}_{N_2}, \mathbf{Y}_1, \dots, \mathbf{Y}_{N_1}\}$ and $\{\mathbf{X}'_1, \dots, \mathbf{X}_{N_2}, \mathbf{Y}_1, \dots, \mathbf{Y}_{N_1}\}$ and denote the respective estimators as $\tilde{\mathbf{G}}_{h_1, h_2}$ and $\tilde{\mathbf{G}}'_{h_1, h_2}$. We have that

$$\begin{aligned} \left| \tilde{\mathbf{G}}_{h_1, h_2} - \tilde{\mathbf{G}}'_{h_1, h_2} \right| &\leq \frac{1}{N_2} \left| g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1) \right) - g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}'_1), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}'_1) \right) \right| \\ &+ \frac{1}{N_2} \sum_{j=2}^{N_2} \left| g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j) \right) - g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}'_{2,h_2}(\mathbf{X}_j) \right) \right|. \end{aligned} \quad (29)$$

Since g is Lipschitz continuous with constant C_g , we have

$$\left| g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1) \right) - g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}'_1), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}'_1) \right) \right| \leq C_g \left(\left| \tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1) - \tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}'_1) \right| + \left| \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1) - \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}'_1) \right| \right) \quad (30)$$

$$\begin{aligned} \left| \tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1) - \tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}'_1) \right| &= \frac{1}{N_1 h_1^d} \left| \sum_{i=1}^{N_1} \left(K \left(\frac{\mathbf{X}_1 - \mathbf{Y}_i}{h_1} \right) - K \left(\frac{\mathbf{X}'_1 - \mathbf{Y}_i}{h_1} \right) \right) \right| \\ &\leq \frac{1}{N_1 h_1^d} \sum_{i=1}^{N_1} \left| K \left(\frac{\mathbf{X}_1 - \mathbf{Y}_i}{h_1} \right) - K \left(\frac{\mathbf{X}'_1 - \mathbf{Y}_i}{h_1} \right) \right| \\ \implies \mathbb{E} \left[\left| \tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1) - \tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}'_1) \right|^2 \right] &\leq \frac{1}{N_1 h_1^{2d}} \sum_{i=1}^{N_1} \mathbb{E} \left[\left(K \left(\frac{\mathbf{X}_1 - \mathbf{Y}_i}{h_1} \right) - K \left(\frac{\mathbf{X}'_1 - \mathbf{Y}_i}{h_1} \right) \right)^2 \right], \end{aligned} \quad (31)$$

where the last step follows from Jensen's inequality. By making the substitution $\mathbf{u}_i = \frac{\mathbf{X}_1 - \mathbf{Y}_i}{h_1}$ and $\mathbf{u}'_i = \frac{\mathbf{X}'_1 - \mathbf{Y}_i}{h_1}$, this gives

$$\begin{aligned} \frac{1}{h_1^{2d}} \mathbb{E} \left[\left(K \left(\frac{\mathbf{X}_1 - \mathbf{Y}_i}{h_1} \right) - K \left(\frac{\mathbf{X}'_1 - \mathbf{Y}_i}{h_1} \right) \right)^2 \right] &= \frac{1}{h^{2d}} \int \left(K \left(\frac{\mathbf{X}_1 - \mathbf{Y}_i}{h_1} \right) - K \left(\frac{\mathbf{X}'_1 - \mathbf{Y}_i}{h_1} \right) \right)^2 \times \\ &\quad f_2(\mathbf{X}_1) f_2(\mathbf{X}'_1) f_1(\mathbf{Y}_i) d\mathbf{X}_1 d\mathbf{X}'_1 d\mathbf{Y}_i \\ &\leq 2 \|K\|_\infty^2. \end{aligned}$$

Combining this with (31) gives

$$\mathbb{E} \left[\left| \tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1) - \tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}'_1) \right|^2 \right] \leq 2 \|K\|_\infty^2.$$

Similarly,

$$\mathbb{E} \left[\left| \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1) - \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}'_1) \right|^2 \right] \leq 2 \|K\|_\infty^2.$$

Combining these results with (30) gives

$$\mathbb{E} \left[\left(g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1) \right) - g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}'_1), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}'_1) \right) \right)^2 \right] \leq 8C_g^2 \|K\|_\infty^2. \quad (32)$$

The second term in (29) is controlled in a similar way. From the Lipschitz condition,

$$\begin{aligned} \left| g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j) \right) - g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}'_{2,h_2}(\mathbf{X}_j) \right) \right|^2 &\leq C_g^2 \left| \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j) - \tilde{\mathbf{f}}'_{2,h_2}(\mathbf{X}_j) \right|^2 \\ &= \frac{C_g^2}{M_2^2 h_2^{2d}} \left(K \left(\frac{\mathbf{X}_j - \mathbf{X}_1}{h} \right) - K \left(\frac{\mathbf{X}_j - \mathbf{X}'_1}{h} \right) \right)^2. \end{aligned}$$

The h_2^{2d} terms are eliminated by making the substitutions of $\mathbf{u}_j = \frac{\mathbf{X}_j - \mathbf{X}_1}{h_2}$ and $\mathbf{u}'_j = \frac{\mathbf{X}_j - \mathbf{X}'_1}{h_2}$ within the expectation to obtain

$$\mathbb{E} \left[\left| g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j) \right) - g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}'_{2,h_2}(\mathbf{X}_j) \right) \right|^2 \right] \leq \frac{2C_g^2 \|K\|_\infty^2}{M_2^2} \quad (33)$$

$$\implies \mathbb{E} \left[\left(\sum_{j=2}^{N_2} \left| g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j) \right) - g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}'_{2,h_2}(\mathbf{X}_j) \right) \right| \right)^2 \right]$$

$$\begin{aligned} &= \sum_{j=2}^{N_2} \sum_{i=2}^{N_2} \mathbb{E} \left[\left| g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j) \right) - g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}'_{2,h_2}(\mathbf{X}_j) \right) \right| \left| g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_i), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_i) \right) - g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_i), \tilde{\mathbf{f}}'_{2,h_2}(\mathbf{X}_i) \right) \right| \right] \\ &\leq M_2^2 \mathbb{E} \left[\left| g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j) \right) - g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}'_{2,h_2}(\mathbf{X}_j) \right) \right|^2 \right] \\ &\leq 2C_g^2 \|K\|_\infty^2, \end{aligned} \quad (34)$$

where we use the Cauchy Schwarz inequality to bound the expectation within each summand. Finally, applying Jensen's inequality and (32) and (34) gives

$$\begin{aligned} \mathbb{E} \left[\left| \tilde{\mathbf{G}}_{h_1,h_2} - \tilde{\mathbf{G}}'_{h_1,h_2} \right|^2 \right] &\leq \frac{2}{N_2^2} \mathbb{E} \left[\left| g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1) \right) - g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}'_1), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}'_1) \right) \right|^2 \right] \\ &\quad + \frac{2}{N_2^2} \mathbb{E} \left[\left(\sum_{j=2}^{N_2} \left| g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j) \right) - g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}'_{2,h_2}(\mathbf{X}_j) \right) \right| \right)^2 \right] \\ &\leq \frac{20C_g^2 \|K\|_\infty^2}{N_2^2}. \end{aligned}$$

Now suppose we have samples $\{\mathbf{X}_1, \dots, \mathbf{X}_{N_2}, \mathbf{Y}_1, \dots, \mathbf{Y}_{N_1}\}$ and $\{\mathbf{X}_1, \dots, \mathbf{X}_{N_2}, \mathbf{Y}'_1, \dots, \mathbf{Y}'_{N_1}\}$ and denote the respective estimators as $\tilde{\mathbf{G}}_{h_1, h_2}$ and $\tilde{\mathbf{G}}'_{h_1, h_2}$. Then

$$\begin{aligned} \left| g\left(\tilde{\mathbf{f}}_{1, h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2, h_2}(\mathbf{X}_j)\right) - g\left(\tilde{\mathbf{f}}'_{1, h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2, h_2}(\mathbf{X}_j)\right) \right| &\leq C_g \left| \tilde{\mathbf{f}}_{1, h_1}(\mathbf{X}_j) - \tilde{\mathbf{f}}'_{1, h_1}(\mathbf{X}_j) \right| \\ &= \frac{C_g}{N_1 h_1^d} \left| K\left(\frac{\mathbf{X}_j - \mathbf{Y}_1}{h_1}\right) - K\left(\frac{\mathbf{X}_j - \mathbf{Y}'_1}{h_1}\right) \right| \\ \implies \mathbb{E} \left[\left| g\left(\tilde{\mathbf{f}}_{1, h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2, h_2}(\mathbf{X}_j)\right) - g\left(\tilde{\mathbf{f}}'_{1, h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2, h_2}(\mathbf{X}_j)\right) \right|^2 \right] &\leq \frac{2C_g^2 \|K\|_\infty^2}{N_1^2}. \end{aligned}$$

Thus using a similar argument as was used to obtain (34),

$$\begin{aligned} \mathbb{E} \left[\left| \tilde{\mathbf{G}}_{h_1, h_2} - \tilde{\mathbf{G}}'_{h_1, h_2} \right|^2 \right] &\leq \frac{1}{N_2^2} \mathbb{E} \left[\left(\sum_{j=1}^{N_2} \left| g\left(\tilde{\mathbf{f}}_{1, h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2, h_2}(\mathbf{X}_j)\right) - g\left(\tilde{\mathbf{f}}'_{1, h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2, h_2}(\mathbf{X}_j)\right) \right| \right)^2 \right] \\ &\leq \frac{2C_g^2 \|K\|_\infty^2}{N_2^2}. \end{aligned}$$

Applying the Efron-Stein inequality gives

$$\mathbb{V} \left[\tilde{\mathbf{G}}_{h_1, h_2} \right] \leq \frac{10C_g^2 \|K\|_\infty^2}{N_2} + \frac{C_g^2 \|K\|_\infty^2 N_1}{N_2^2}.$$

APPENDIX D PROOF OF THEOREM 5

We are interested in the asymptotic distribution of

$$\begin{aligned} \sqrt{N_2} \left(\tilde{\mathbf{G}}_{h_1, h_2} - \mathbb{E} \left[\tilde{\mathbf{G}}_{h_1, h_2} \right] \right) &= \frac{1}{\sqrt{N_2}} \sum_{j=1}^{N_2} \left(g\left(\tilde{\mathbf{f}}_{1, h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2, h_2}(\mathbf{X}_j)\right) - \mathbb{E}_{\mathbf{X}_j} \left[g\left(\tilde{\mathbf{f}}_{1, h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2, h_2}(\mathbf{X}_j)\right) \right] \right) \\ &\quad + \frac{1}{\sqrt{N_2}} \sum_{j=1}^{N_2} \left(\mathbb{E}_{\mathbf{X}_j} \left[g\left(\tilde{\mathbf{f}}_{1, h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2, h_2}(\mathbf{X}_j)\right) \right] - \mathbb{E} \left[g\left(\tilde{\mathbf{f}}_{1, h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2, h_2}(\mathbf{X}_j)\right) \right] \right). \end{aligned}$$

Note that by the standard central limit theorem [59], the second term converges in distribution to a Gaussian random variable. If the first term converges in probability to a constant (specifically, 0), then we can use Slutsky's theorem [60] to find the asymptotic distribution. So now we focus on the first term which we denote as \mathbf{V}_{N_2} .

To prove convergence in probability, we will use Chebyshev's inequality. Note that $\mathbb{E}[\mathbf{V}_{N_2}] = 0$. To bound the variance of \mathbf{V}_{N_2} , we again use the Efron-Stein inequality. Let \mathbf{X}'_1 be drawn from f_2 and denote \mathbf{V}_{N_2} and \mathbf{V}'_{N_2} as the sequences using \mathbf{X}_1 and \mathbf{X}'_1 , respectively. Then

$$\begin{aligned} \mathbf{V}_{N_2} - \mathbf{V}'_{N_2} &= \frac{1}{\sqrt{N_2}} \left(g\left(\tilde{\mathbf{f}}_{1, h_1}(\mathbf{X}_1), \tilde{\mathbf{f}}_{2, h_2}(\mathbf{X}_1)\right) - \mathbb{E}_{\mathbf{X}_1} \left[g\left(\tilde{\mathbf{f}}_{1, h_1}(\mathbf{X}_1), \tilde{\mathbf{f}}_{2, h_2}(\mathbf{X}_1)\right) \right] \right) \\ &\quad + \frac{1}{\sqrt{N_2}} \left(g\left(\tilde{\mathbf{f}}_{1, h_1}(\mathbf{X}'_1), \tilde{\mathbf{f}}_{2, h_2}(\mathbf{X}'_1)\right) - \mathbb{E}_{\mathbf{X}'_1} \left[g\left(\tilde{\mathbf{f}}_{1, h_1}(\mathbf{X}'_1), \tilde{\mathbf{f}}_{2, h_2}(\mathbf{X}'_1)\right) \right] \right) \\ &\quad + \frac{1}{\sqrt{N_2}} \sum_{j=2}^{N_2} \left(g\left(\tilde{\mathbf{f}}_{1, h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2, h_2}(\mathbf{X}_j)\right) - g\left(\tilde{\mathbf{f}}_{1, h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}'_{2, h_2}(\mathbf{X}_j)\right) \right). \end{aligned} \tag{35}$$

Note that

$$\mathbb{E} \left[\left(g\left(\tilde{\mathbf{f}}_{1, h_1}(\mathbf{X}_1), \tilde{\mathbf{f}}_{2, h_2}(\mathbf{X}_1)\right) - \mathbb{E}_{\mathbf{X}_1} \left[g\left(\tilde{\mathbf{f}}_{1, h_1}(\mathbf{X}_1), \tilde{\mathbf{f}}_{2, h_2}(\mathbf{X}_1)\right) \right] \right)^2 \right] = \mathbb{E} \left[\mathbb{V}_{\mathbf{X}_1} \left[g\left(\tilde{\mathbf{f}}_{1, h_1}(\mathbf{X}_1), \tilde{\mathbf{f}}_{2, h_2}(\mathbf{X}_1)\right) \right] \right].$$

If we condition on \mathbf{X}_1 , then by the standard central limit theorem $\sqrt{N_i h_i^d} \left(\tilde{\mathbf{f}}_{i, h_i}(\mathbf{X}_1) - \mathbb{E}_{\mathbf{X}_1} \left[\tilde{\mathbf{f}}_{i, h_i}(\mathbf{X}_1) \right] \right)$ converges in distribution to a zero mean Gaussian random variable with variance $\sigma_{\tilde{\mathbf{f}}_i}^2(\mathbf{X}_1) = O(1)$. This is true even if \mathbf{X}_1 is close to the boundary of the support of the densities. The KDEs $\tilde{\mathbf{f}}_{1, h_1}(\mathbf{X}_1)$ and $\tilde{\mathbf{f}}_{2, h_2}(\mathbf{X}_1)$ are conditionally independent given \mathbf{X}_1 as are their limiting distributions. Thus the KDEs converge jointly in distribution to a Gaussian random vector with zero mean,

zero covariance, and their respective variances. By the delta method [61], we have that if $g(x, y)$ is continuously differentiable with respect to both x and y at $\mathbb{E}_{\mathbf{X}_1} [\tilde{\mathbf{f}}_{i,h_i}(\mathbf{X}_1)]$ for $i = 1, 2$, respectively, then

$$\mathbb{V}_{\mathbf{X}_1} \left[g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1) \right) \right] = O \left(\frac{1}{N_1 h_1^d} + \frac{1}{N_2 h_2^d} \right) = o(1),$$

provided that $N_i h_i^d \rightarrow \infty$. Thus $\mathbb{E} \left[\mathbb{V}_{\mathbf{X}_1} \left[g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1) \right) \right] \right] = o(1)$. A similar result holds when we replace \mathbf{X}_1 with \mathbf{X}'_1 .

For the third term in (35),

$$\begin{aligned} & \mathbb{E} \left[\left(\sum_{j=2}^{N_2} \left| g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j) \right) - g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}'_{2,h_2}(\mathbf{X}_j) \right) \right| \right)^2 \right] \\ &= \sum_{j=2}^{N_2} \sum_{i=2}^{N_2} \mathbb{E} \left[\left| g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j) \right) - g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}'_{2,h_2}(\mathbf{X}_j) \right) \right| \left| g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_i), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_i) \right) - g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_i), \tilde{\mathbf{f}}'_{2,h_2}(\mathbf{X}_i) \right) \right| \right]. \end{aligned}$$

There are M_2 terms where $i = j$ and we have from Appendix C (see (33)) that

$$\mathbb{E} \left[\left| g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j) \right) - g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}'_{2,h_2}(\mathbf{X}_j) \right) \right|^2 \right] \leq \frac{2C_g^2 \|K\|_\infty^2}{M_2^2}.$$

Thus these terms are $O\left(\frac{1}{M_2}\right)$. There are $M_2^2 - M_2$ terms when $i \neq j$. In this case, we can do four substitutions of the form $\mathbf{u}_j = \frac{\mathbf{X}_j - \mathbf{X}_1}{h_2}$ to obtain

$$\mathbb{E} \left[\left| g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j) \right) - g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}'_{2,h_2}(\mathbf{X}_j) \right) \right| \left| g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_i), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_i) \right) - g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_i), \tilde{\mathbf{f}}'_{2,h_2}(\mathbf{X}_i) \right) \right| \right] \leq \frac{4C_g^2 \|K\|_\infty^2 h_2^{2d}}{M_2^2}.$$

Then since $h_2^d = o(1)$, we get

$$\mathbb{E} \left[\left(\sum_{j=2}^{N_2} \left| g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j) \right) - g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}'_{2,h_2}(\mathbf{X}_j) \right) \right| \right)^2 \right] = o(1), \quad (36)$$

$$\begin{aligned} \implies \mathbb{E} \left[\left(\mathbf{V}_{N_2} - \mathbf{V}'_{N_2} \right)^2 \right] &\leq \frac{3}{N_2} \mathbb{E} \left[\left(g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1) \right) - \mathbb{E}_{\mathbf{X}_1} \left[g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1) \right) \right] \right)^2 \right] \\ &\quad + \frac{3}{N_2} \mathbb{E} \left[\left(g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}'_1), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}'_1) \right) - \mathbb{E}_{\mathbf{X}'_1} \left[g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}'_1), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}'_1) \right) \right] \right)^2 \right] \\ &\quad + \frac{3}{N_2} \mathbb{E} \left[\left(\sum_{j=2}^{N_2} \left(g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j) \right) - g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}'_{2,h_2}(\mathbf{X}_j) \right) \right) \right)^2 \right] \\ &= o\left(\frac{1}{N_2}\right). \end{aligned}$$

Now consider samples $\{\mathbf{X}_1, \dots, \mathbf{X}_{N_2}, \mathbf{Y}_1, \dots, \mathbf{Y}_{N_1}\}$ and $\{\mathbf{X}_1, \dots, \mathbf{X}_{N_2}, \mathbf{Y}'_1, \dots, \mathbf{Y}_{N_1}\}$ and the respective sequences \mathbf{V}_{N_2} and \mathbf{V}'_{N_2} . Then

$$\mathbf{V}_{N_2} - \mathbf{V}'_{N_2} = \frac{1}{\sqrt{N_2}} \sum_{j=1}^{N_2} \left(g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j) \right) - g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}'_{2,h_2}(\mathbf{X}_j) \right) \right).$$

Using a similar argument as that used to obtain (36), we have that if $h_1^d = o(1)$ and $N_1 \rightarrow \infty$, then

$$\mathbb{E} \left[\left(\sum_{j=2}^{N_2} \left| g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j) \right) - g \left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}'_{2,h_2}(\mathbf{X}_j) \right) \right| \right)^2 \right] = o(1)$$

$$\implies \mathbb{E} \left[\left(\mathbf{V}_{N_2} - \mathbf{V}'_{N_2} \right)^2 \right] = o \left(\frac{1}{N_2} \right).$$

Applying the Efron-Stein inequality gives

$$\mathbb{V} [\mathbf{V}_{N_2}] = o \left(\frac{N_2 + N_1}{N_2} \right) = o(1).$$

Thus by Chebyshev's inequality,

$$\Pr (|\mathbf{V}_{N_2}| > \epsilon) \leq \frac{\mathbb{V} [\mathbf{V}_{N_2}]}{\epsilon^2} = o(1),$$

and therefore \mathbf{V}_{N_2} converges to zero in probability. By Slutsky's theorem, $\sqrt{N_2} \left(\tilde{\mathbf{G}}_{h_1, h_2} - \mathbb{E} [\tilde{\mathbf{G}}_{h_1, h_2}] \right)$ converges in distribution to a zero mean Gaussian random variable with variance

$$\mathbb{V} \left[\mathbb{E}_{\mathbf{X}} \left[g \left(\tilde{\mathbf{f}}_{1, h_1}(\mathbf{X}), \tilde{\mathbf{f}}_{2, h_2}(\mathbf{X}) \right) \right] \right],$$

where \mathbf{X} is drawn from f_2 .

For the weighted ensemble estimator, we wish to know the asymptotic distribution of $\sqrt{N_2} \left(\tilde{\mathbf{G}}_w - \mathbb{E} [\tilde{\mathbf{G}}_w] \right)$ where $\tilde{\mathbf{G}}_w = \sum_{l \in \bar{l}} w(l) \tilde{\mathbf{G}}_{h(l)}$. We have that

$$\begin{aligned} \sqrt{N_2} \left(\tilde{\mathbf{G}}_w - \mathbb{E} [\tilde{\mathbf{G}}_w] \right) &= \frac{1}{\sqrt{N_2}} \sum_{j=1}^{N_2} \sum_{l \in \bar{l}} w(l) \left(g \left(\tilde{\mathbf{f}}_{1, h(l)}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2, h(l)}(\mathbf{X}_j) \right) - \mathbb{E}_{\mathbf{X}_j} \left[g \left(\tilde{\mathbf{f}}_{1, h(l)}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2, h(l)}(\mathbf{X}_j) \right) \right] \right) \\ &\quad + \frac{1}{\sqrt{N_2}} \sum_{j=1}^{N_2} \left(\mathbb{E}_{\mathbf{X}_j} \left[\sum_{l \in \bar{l}} w(l) g \left(\tilde{\mathbf{f}}_{1, h(l)}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2, h(l)}(\mathbf{X}_j) \right) \right] - \mathbb{E} \left[\sum_{l \in \bar{l}} w(l) g \left(\tilde{\mathbf{f}}_{1, h(l)}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2, h(l)}(\mathbf{X}_j) \right) \right] \right). \end{aligned}$$

The second term again converges in distribution to a Gaussian random variable by the central limit theorem. The mean and variance are, respectively, zero and

$$\mathbb{V} \left[\sum_{l \in \bar{l}} w(l) \mathbb{E}_{\mathbf{X}} \left[g \left(\tilde{\mathbf{f}}_{1, h(l)}(\mathbf{X}), \tilde{\mathbf{f}}_{2, h(l)}(\mathbf{X}) \right) \right] \right].$$

The first term is equal to

$$\begin{aligned} \sum_{l \in \bar{l}} w(l) \left(\frac{1}{\sqrt{N_2}} \sum_{j=1}^{N_2} \left(g \left(\tilde{\mathbf{f}}_{1, h(l)}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2, h(l)}(\mathbf{X}_j) \right) - \mathbb{E}_{\mathbf{X}_j} \left[g \left(\tilde{\mathbf{f}}_{1, h(l)}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2, h(l)}(\mathbf{X}_j) \right) \right] \right) \right) &= \sum_{l \in \bar{l}} w(l) o_P(1) \\ &= o_P(1), \end{aligned}$$

where $o_P(1)$ denotes convergence to zero in probability. In the last step, we used the fact that if two random variables converge in probability to constants, then their linear combination converges in probability to the linear combination of the constants. Combining this result with Slutsky's theorem completes the proof.

REFERENCES

- [1] T. M. Cover and J. A. Thomas, *Elements of information theory*, John Wiley & Sons, 2012.
- [2] H. Avi-Itzhak and T. Diep, "Arbitrarily tight upper and lower bounds on the Bayesian probability of error," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 1, pp. 89–91, 1996.
- [3] W. A. Hashlamoun, P. K. Varshney, and V. Samarasekera, "A tight upper bound on the Bayesian probability of error," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 2, pp. 220–224, 1994.
- [4] K.R. Moon, V. Delouille, and A. O. Hero III, "Meta learning of bounds on the Bayes classifier error," in *IEEE Signal Processing and SP Education Workshop*. IEEE, 2015, pp. 13–18.
- [5] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *The Annals of Mathematical Statistics*, pp. 493–507, 1952.
- [6] V. Berisha, A. Wisler, A. O. Hero III, and A. Spanias, "Empirically estimable classification bounds based on a new divergence measure," *IEEE Transactions on Signal Processing*, 2015.
- [7] K. R. Moon and A. O. Hero III, "Multivariate f-divergence estimation with confidence," in *Advances in Neural Information Processing Systems*, 2014, pp. 2420–2428.
- [8] Stephen V Gliske, Kevin R Moon, William C Stacey, and Alfred O Hero III, "The intrinsic value of HFO features as a biomarker of epileptic activity," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2016.
- [9] B. Póczos and J. G. Schneider, "On the estimation of alpha-divergences," in *International Conference on Artificial Intelligence and Statistics*, 2011, pp. 609–617.
- [10] J. Oliva, B. Póczos, and J. Schneider, "Distribution to distribution regression," in *Proceedings of The 30th International Conference on Machine Learning*, 2013, pp. 1049–1057.
- [11] Z. Szabó, A. Gretton, B. Póczos, and B. Sriperumbudur, "Two-stage sampled learning theory on distributions," *To appear in AISTATS*, 2015.

- [12] K. R. Moon, V. Delouille, J. J. Li, R. De Visscher, F. Watson, and A. O. Hero III, "Image patch analysis of sunspots and active regions. II. Clustering via matrix factorization," *Journal of Space Weather and Space Climate*, vol. 6, no. A3, 2016.
- [13] K. R. Moon, J. J. Li, V. Delouille, R. De Visscher, F. Watson, and A. O. Hero III, "Image patch analysis of sunspots and active regions. I. Intrinsic dimension and correlation analysis," *Journal of Space Weather and Space Climate*, vol. 6, no. A2, 2016.
- [14] I. S. Dhillon, S. Mallela, and R. Kumar, "A divisive information theoretic feature clustering algorithm for text classification," *The Journal of Machine Learning Research*, vol. 3, pp. 1265–1287, 2003.
- [15] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *The Journal of Machine Learning Research*, vol. 6, pp. 1705–1749, 2005.
- [16] J. Lewi, R. Butera, and L. Paninski, "Real-time adaptive information-theoretic optimization of neurophysiology experiments," in *Advances in Neural Information Processing Systems*, 2006, pp. 857–864.
- [17] L. Bruzzone, F. Roli, and S. B. Serpico, "An extension of the Jeffreys-Matusita distance to multiclass cases for feature selection," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 33, no. 6, pp. 1318–1321, 1995.
- [18] X. Guorong, C. Peiqi, and W. Minhui, "Bhattacharyya distance feature selection," in *Pattern Recognition, 1996., Proceedings of the 13th International Conference on*. IEEE, 1996, vol. 2, pp. 195–199.
- [19] D. M. Sakate and D. N. Kashid, "Variable selection via penalized minimum φ -divergence estimation in logistic regression," *Journal of Applied Statistics*, vol. 41, no. 6, pp. 1233–1246, 2014.
- [20] K. E. Hild, D. Erdogmus, and J. C. Principe, "Blind source separation using Renyi's mutual information," *Signal Processing Letters, IEEE*, vol. 8, no. 6, pp. 174–176, 2001.
- [21] M. Mihoko and S. Eguchi, "Robust blind source separation by beta divergence," *Neural computation*, vol. 14, no. 8, pp. 1859–1886, 2002.
- [22] B. C. Vemuri, M. Liu, S. Amari, and F. Nielsen, "Total Bregman divergence and its applications to DTI analysis," *Medical Imaging, IEEE Transactions on*, vol. 30, no. 2, pp. 475–483, 2011.
- [23] A. B. Hamza and H. Krim, "Image registration and segmentation by maximizing the Jensen-Rényi divergence," in *Energy Minimization Methods in Computer Vision and Pattern Recognition*. Springer, 2003, pp. 147–163.
- [24] G. Liu, G. Xia, W. Yang, and N. Xue, "SAR image segmentation via non-local active contours," in *Geoscience and Remote Sensing Symposium (IGARSS), 2014 IEEE International*. IEEE, 2014, pp. 3730–3733.
- [25] V. Korzhik and I. Fedyanin, "Steganographic applications of the nearest-neighbor approach to Kullback-Leibler divergence estimation," in *Digital Information, Networking, and Wireless Communications (DINWC), 2015 Third International Conference on*. IEEE, 2015, pp. 133–138.
- [26] M. Basseville, "Divergence measures for statistical data processing—An annotated bibliography," *Signal Processing*, vol. 93, no. 4, pp. 621–633, 2013.
- [27] B. Chai, D. Walther, D. Beck, and L. Fei-Fei, "Exploring functional connectivities of the human brain using multivariate information analysis," in *Advances in neural information processing systems*, 2009, pp. 270–278.
- [28] K. M. Carter, R. Raich, and A. O. Hero III, "On local intrinsic dimension estimation and its applications," *Signal Processing, IEEE Transactions on*, vol. 58, no. 2, pp. 650–663, 2010.
- [29] K. R. Moon, J. J. Li, V. Delouille, F. Watson, and A. O. Hero III, "Image patch analysis and clustering of sunspots: A dimensionality reduction approach," in *IEEE International Conference on Image Processing*. IEEE, 2014, pp. 1623–1627.
- [30] A. O. Hero III, B. Ma, O. Michel, and J. Gorman, "Applications of entropic spanning graphs," *Signal Processing Magazine, IEEE*, vol. 19, no. 5, pp. 85–95, 2002.
- [31] I. Csiszar, "Information-type measures of difference of probability distributions and indirect observations," *Studia Sci. Math. Hungar.*, vol. 2, pp. 299–318, 1967.
- [32] S. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 131–142, 1966.
- [33] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.
- [34] A. Rényi, "On measures of entropy and information," in *Fourth Berkeley Sympos. on Mathematical Statistics and Probability*, 1961, pp. 547–561.
- [35] E. Hellinger, "Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen," *Journal für die reine und angewandte Mathematik*, vol. 136, pp. 210–271, 1909.
- [36] A. Bhattacharyya, "On a measure of divergence between two multinomial populations," *Sankhyā: The Indian Journal of Statistics*, pp. 401–406, 1946.
- [37] K. Sricharan, D. Wei, and A. O. Hero, "Ensemble estimators for multivariate entropy estimation," *Information Theory, IEEE Transactions on*, vol. 59, no. 7, pp. 4374–4388, 2013.
- [38] Qing Wang, Sanjeev R Kulkarni, and Sergio Verdú, "Divergence estimation for multidimensional densities via k-nearest-neighbor distances," *IEEE Trans. Information Theory*, vol. 55, no. 5, pp. 2392–2405, 2009.
- [39] Georges A Darbellay, Igor Vajda, et al., "Estimation of the information by an adaptive partitioning of the observation space," *IEEE Trans. Information Theory*, vol. 45, no. 4, pp. 1315–1321, 1999.
- [40] Jorge Silva and Shrikanth S Narayanan, "Information divergence estimation based on data-dependent partitions," *Journal of Statistical Planning and Inference*, vol. 140, no. 11, pp. 3180–3198, 2010.
- [41] Trung Kien Le, "Information dependency: Strong consistency of Darbellay–Vajda partition estimators," *Journal of Statistical Planning and Inference*, vol. 143, no. 12, pp. 2089–2100, 2013.
- [42] Qing Wang, Sanjeev R Kulkarni, and Sergio Verdú, "Divergence estimation of continuous distributions based on data-dependent partitions," *IEEE Trans. Information Theory*, vol. 51, no. 9, pp. 3064–3074, 2005.
- [43] K. R. Moon and A. O. Hero III, "Ensemble estimation of multivariate f-divergence," in *Information Theory (ISIT), 2014 IEEE International Symposium on*. IEEE, 2014, pp. 356–360.
- [44] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "Estimating divergence functionals and the likelihood ratio by convex risk minimization," *Information Theory, IEEE Transactions on*, vol. 56, no. 11, pp. 5847–5861, 2010.
- [45] A. Krishnamurthy, K. Kandasamy, B. Poczos, and L. Wasserman, "Nonparametric estimation of renyi divergence and friends," in *Proceedings of The 31st International Conference on Machine Learning*, 2014, pp. 919–927.
- [46] S. Singh and B. Póczos, "Generalized exponential concentration inequality for rényi divergence estimation," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 2014, pp. 333–341.
- [47] S. Singh and B. Póczos, "Exponential concentration of a density functional estimator," in *Advances in Neural Information Processing Systems*, 2014, pp. 3032–3040.
- [48] Kirthevasan Kandasamy, Akshay Krishnamurthy, Barnabas Poczos, Larry Wasserman, and James Robins, "Nonparametric von mises estimators for entropies, divergences and mutual informations," in *Advances in Neural Information Processing Systems*, 2015, pp. 397–405.
- [49] Wolfgang Härdle, *Applied Nonparametric Regression*, Cambridge University Press, 1990.
- [50] A. Berline, L. Devroye, and L. Györfi, "Asymptotic normality of L1 error in density estimation," *Statistics*, vol. 26, pp. 329–343, 1995.
- [51] A. Berline, László Györfi, and István Dénes, "Asymptotic normality of relative entropy in multivariate density estimation," *Publications de l'Institut de Statistique de l'Université de Paris*, vol. 41, pp. 3–27, 1997.

- [52] Peter J Bickel and Murray Rosenblatt, "On some global measures of the deviations of density function estimates," *The Annals of Statistics*, pp. 1071–1095, 1973.
- [53] Lucien Birgé and Pascal Massart, "Estimation of integral functionals of a density," *The Annals of Statistics*, pp. 11–29, 1995.
- [54] Evarist Giné and David M Mason, "Uniform in bandwidth estimation of integral functionals of the density function," *Scandinavian Journal of Statistics*, vol. 35, no. 4, pp. 739–761, 2008.
- [55] Béatrice Laurent et al., "Efficient estimation of integral functionals of a density," *The Annals of Statistics*, vol. 24, no. 2, pp. 659–681, 1996.
- [56] Kumar Sricharan, Raviv Raich, and Alfred O Hero, "Estimation of nonlinear functionals of densities with confidence," *IEEE Trans. Information Theory*, vol. 58, no. 7, pp. 4135–4159, 2012.
- [57] Bruce E Hansen, "Lecture notes on nonparametrics," 2009.
- [58] Bradley Efron and Charles Stein, "The jackknife estimate of variance," *The Annals of Statistics*, pp. 586–596, 1981.
- [59] Rick Durrett, *Probability: Theory and Examples*, Cambridge University Press, 2010.
- [60] Allan Gut, *Probability: A Graduate Course*, Springer Science & Business Media, 2012.
- [61] Robert Keener, *Theoretical Statistics: Topics for a Core Course*, Springer Science & Business Media, 2010.